Statistical Soil-landscape Modelling for Environmental Management

by
Paul Edward Gessler

March, 1996

A thesis submitted for the degree of
Doctor of Philosophy of The Australian National University

Statement

I hereby state that the work described and contained within this thesis is my own, unless expressly stated otherwise in the text.

Paul Edward Gessler

Paul Edward Gessler
March, 1996

Acknowledgements

Ian Moore and I had many fruitful and synergistic discussions that kicked off the ideas explored in this thesis. I gratefully acknowledge the time I shared with him and will forever mourn the loss of a great man taken too early. Continued discussions along the way with Neil McKenzie, Mike Hutchinson, Chris Moran, Henry Nix, John Gallant, Kevin McSweeney, Jay Bell, and Jerry Nielsen assisted the progression of ideas. They all deserve acknowledgement and thanks.

Neil McKenzie has been a supportive colleague in his many roles of supervisor, project leader, friend and punching bag. He always kept an even keel. The CSIRO Division of Soils is thankfully acknowledged for the resource support and facilities that made this work possible. Hutch, Henry and staff at the Centre for Resource and Environmental Studies provided an administrative home for this PhD. Partial funding was provided by grant M218, "Developing spatial analysis methods for land resource assessment", from the Murray-Darling Basin Commission.

Many others helped at various times including John Hutka, Dermot McKane, Deborah O'Connell, Jim Beatty, Jeff Wood, Tony Butler, John Williams, Colin Chartres, Jenny Kesteven and Phil Bierwirth. Linda Ashton has been an assistant extraordinaire who absorbed what I could explain and then began to find the best path on her own (even if she couldn't spot the pegs in the field!). She deserves a special thanks. On numerous occasions, Chris Moran helped me transform my inefficient code into something that whistled "Waltzing Matilda" - thanks.

This thesis is dedicated to my wife, Kathy. She left behind everything familiar to venture to Australia with me. She always believed that my (often distant) promise to spend more time with family would eventuate. Words won't do justice to my gratitude and love. May this work contribute to more informed management and understanding of the natural environment that my children, Travis and Adriana, will inherit from us. Thanks to Travis for always demanding that my attention focus onto something fun and simple, and to Adriana for that incessant smile and twinkle in her eye.

Abstract

Soil-landscape modelling and the production of soil maps have traditionally relied on a sampling of the soil continuum to establish environmental correlations useful for spatial extension and extrapolation. Seldom are the correlations explicitly formalized and quantified so that they may be repeated, tested or improved. This work develops statistical soil-landscape models of individual soil attributes using quantitative correlations between soil attribute measurements and digital environmental variables. These environmental variables include high resolution digital elevation models (DEM's), airborne gamma radiometric imagery and climatic surfaces. The models are used to generate spatial predictions of soil attributes, visualize soil patterns and evaluate hypotheses regarding landscape process interpretations and development of quantitative hillslope models representative of three-dimensional landscapes in the study areas.

A provisional model is used to define a terrain attribute environmental gradient for a detailed stratified random field sampling of three study areas in southeastern Australia. Exploratory data analysis and confirmatory statistical models are developed using generalized linear, generalized additive, regression and classification Tree models. The models are implemented for spatial prediction using continuous environmental variables and map algebra within a geographical information system. Results indicate that a broad range of modelling tools and explanatory environmental attributes are required depending on the variation patterns exhibited by the response soil attribute sample set.

At local hillslope scales, response soil attributes exhibited strong correlations with digital terrain attributes computed from 20m digital elevation models. Evaluation of DEM resolution indicates that useful correlations exist over several scales (5m-40m grid point spacings) but decline markedly at a grid point spacing of 80m, the largest grid point spacing evaluated. Spatial analysis tools enabled the development and visualization of spatially-averaged convergent and divergent hillslope models for each study area. These techniques provide a useful framework for developing hypotheses about landscape processes for further testing and improved understanding.

Table of Contents

List of Figures

List of Tables

Chapter One: Introduction

1.1 BROAD PRINCIPLES

Soil occupies the interface in the terrestrial biosphere between the lithosphere and atmosphere and modifies many material and energy exchanges, pathways and cycles. Earth surface processes operate over a variety of spatio-temporal scales often generating complex soil patterns that are difficult to model and visualize in space. The hidden or underfoot nature of the soil mantle further complicates the modelling of soil spatial patterns. Therefore, correlations between finite samples of the soil mantle and more easily observed patterns of landform, vegetation and other environmental variables have been the principle means for mapping and extrapolating soil patterns by conventional soil survey at local hillslope scales. This process is inherently statistical in that samples are used to infer properties of a population.

However, the theory supporting field sampling, data analysis, environmental correlation and spatial prediction in conventional soil survey has received inadequate attention and current methods tend to be qualitative and ill-defined (Butler, 1964; 1980; Burrough, 1986; Holmgren, 1988; Hudson, 1992; Bell *et al.*, 1992; Moore *et al.*, 1993; McKenzie and Austin, 1993; Hewitt, 1993; McSweeney *et al.*, 1994). Soil survey can be improved by making methods more explicit and repeatable (Hudson, 1992; Hewitt, 1993). One approach to achieve this is by using computer-based spatial modelling to develop explicit and quantitative environmental correlations and statistical soil-landscape models for predicting soil attribute distribution. The intention of the approach is not to eschew conventional methods, but to build on them using contemporary tools (e.g. GIS, statistical software, digital environmental data) that preserve primary data. This retains the capacity for re-analysis and improvement of predictions as methods evolve and new data become available.

McKenzie and Austin (1993) and McSweeney *et al.* (1994) outlined new conceptual approaches for modelling the soil-landscape continuum, but did not provide examples or specific details of the mechanics for broad spatial implementation. The basic aim of this thesis is to carry these concepts forward by demonstrating an integration of new tools for development and application of statistical models of soil attributes and patterns.

1.2 DEFINITIONS

Before outlining the thesis structure, key terms and concepts require definition. These will be expanded on throughout the thesis.

1.2.1 Soil-landscape Model

Models describe or imitate the properties of other "real" objects in a simpler or more convenient form often using a descriptive language (Chambers and Hastie, 1992). The term soil-landscape has been used and defined in many ways (Jenny, 1941; Mabbutt, 1968; Huggett, 1975; Thompson and Moore, 1984; Northcote, 1984; Hole and Campbell, 1985; Hudson, 1992; Dobrovol'skiy, 1994). Soil-landscape models are a central feature in current soil survey. Slater *et al.* (1994) define soil-landscape models as "expressions of the linkage of geosphere and biosphere components as products of historical process." Defined in a complementary way, soil-landscape models relate soil patterns, usually determined by field sampling, to the landscape and environmental context (e.g. climate, geology, landform, vegetation, land use).

The models may range from: verbal descriptions of soil patterns using an ethnic language (Pawluck *et al.*, 1992; Zimmerer, 1994); to descriptions using points, lines and polygons (e.g. soil survey and land system maps) using an integration of taxonomic and cartographic languages (Christian and Stewart, 1968; Dent and Young, 1981; Soil Survey Division Staff, 1993); to mathematical descriptions of soil horizon or specific soil attribute patterns using environmental correlations and a

statistical language (Shovic and Montagne, 1985; Bouma, 1989; Bell *et al.* 1992, 1994; Moore *et al.* 1993; McKenzie and Austin, 1993; Odeh *et al.* 1994). Soil-landscape models may be used to:

- document and transfer experience or knowledge;
- develop explicit sampling schemes;
- predict soil attributes;
- explore and compare attributes related to spatial soil processes;
- develop and test hypotheses; and
- provide inputs to broader environmental and socioeconomic models (Hudson, 1992; Moore *et al.* 1993; Hewitt, 1993).

In this thesis, soil-landscape models will be represented by statistical models that have environmental variables as predictors and soil attributes as response variables. The soil-landscape models are semi-empirical and have variables that reflect hypothesized processes of pedogenesis. A critical issue in the development of these models is the scale of measurement or observation.

1.2.2 Scale and Measurement of Environmental Variables

Terms for scale and measurement are defined in various ways by different disciplines. Figure 1.1 displays a simple model and nomenclature of the space-time continuum based on the SI units ***metre*** and ***second***. References to scale and measurement will use this nomenclature throughout the thesis.

Environmental variables are measured or predicted over the earth at discrete or aggregated intervals of the space-time continuum that may be broadly grouped as climatic, geologic, geophysical, topographic, soil and others. Environmental variables may be measured using different attribute types. Austin and McKenzie (1988) differentiate four overlapping groups as:

- nominal attributes - discrete classification (e.g. presence versus absence, bedrock geology type, soil map unit);

- ordinal attributes - discrete classes with order (e.g. frequency of inundation, abundance of soil mottles, horizon type);
- interval attributes - measured on a continuous scale with no true zero (e.g. pH, temperature in °C);
- ratio scale attributes - measured on a continuous scale with a true zero (e.g. solum depth, total carbon).

They also identify *serially dependent attributes* that are dependent on the presence of another (e.g. E horizon depth if E horizons are present) and *profile attributes* where several measurements of the same variable are a linked set (e.g. total carbon measured at depth increments down the soil profile). Some environmental variables are easy and inexpensive to measure in a continuous manner over broad spatial areas (e.g. slope from digital elevation models), while others are time consuming and expensive to measure on small samples (e.g. soil hydraulic conductivity of a 0.2m core from an individual soil horizon).

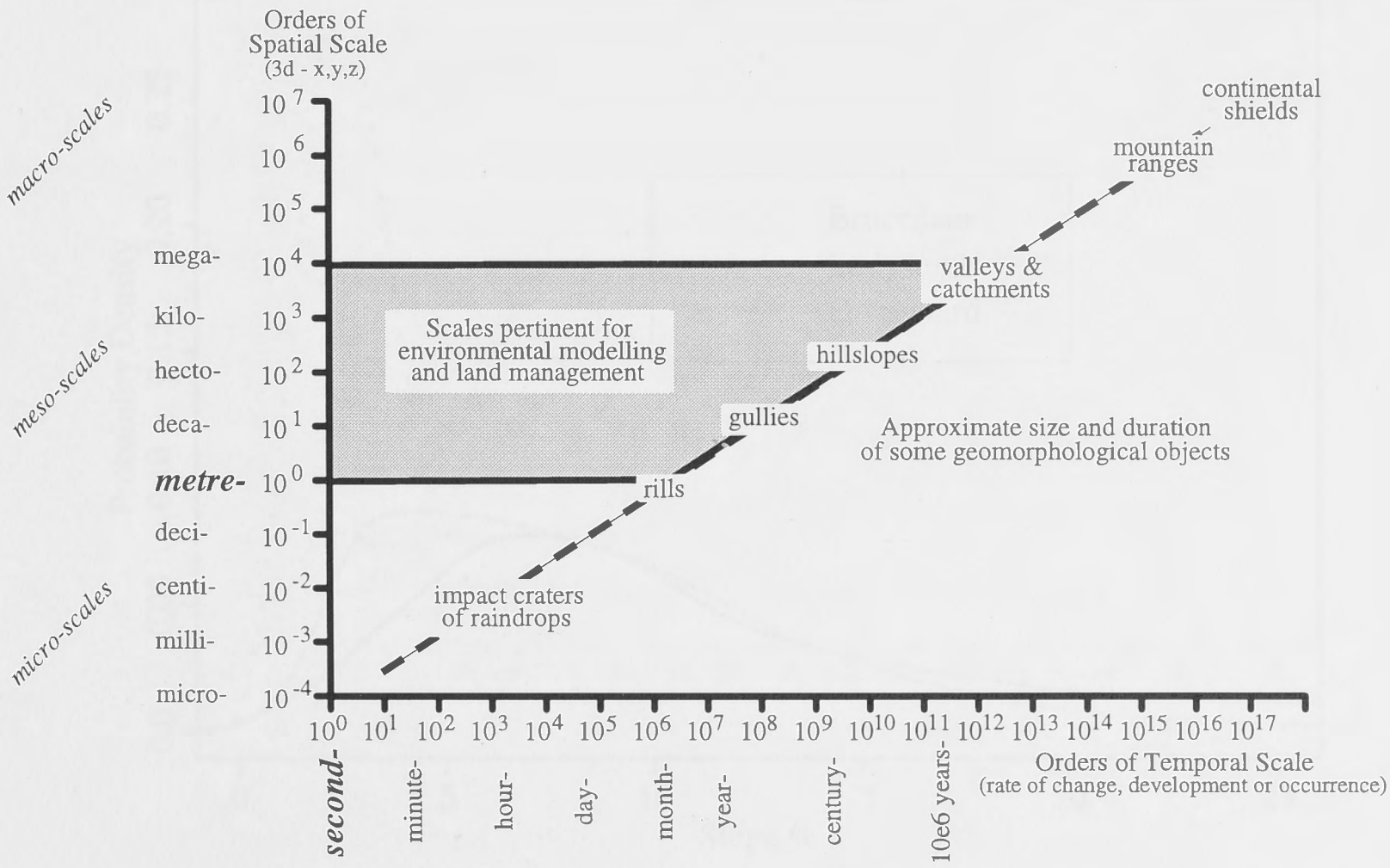


Figure 1.1 Space-Time Continuum (modified from Ahnert, 1988; Dikau, 1989)

Two important concepts that build on these definitions are *environmental attribute* and *geographic* spaces. Figure 1.2 displays the relative probability density functions or attribute distributions for the slope gradient (%) environmental variable measured for each 20m x 20m digital elevation model (DEM) grid cell over the geographic extent of the three study areas used in this thesis (see Figure 3.1). Because these lower meso-scale slope measurements are available in a spatially continuous manner over each geographic space, they may be considered models of the slope populations for each area. Thus, they are indicative of the underlying generative probability process(es). Process interpretations are dependent on informed stratification and exploration of variation in the area under study. Environmental correlation is the quantitative definition of relationships between distributions or samples from one, both or several distributions. These concepts will be expanded on in Chapters Three and Four.

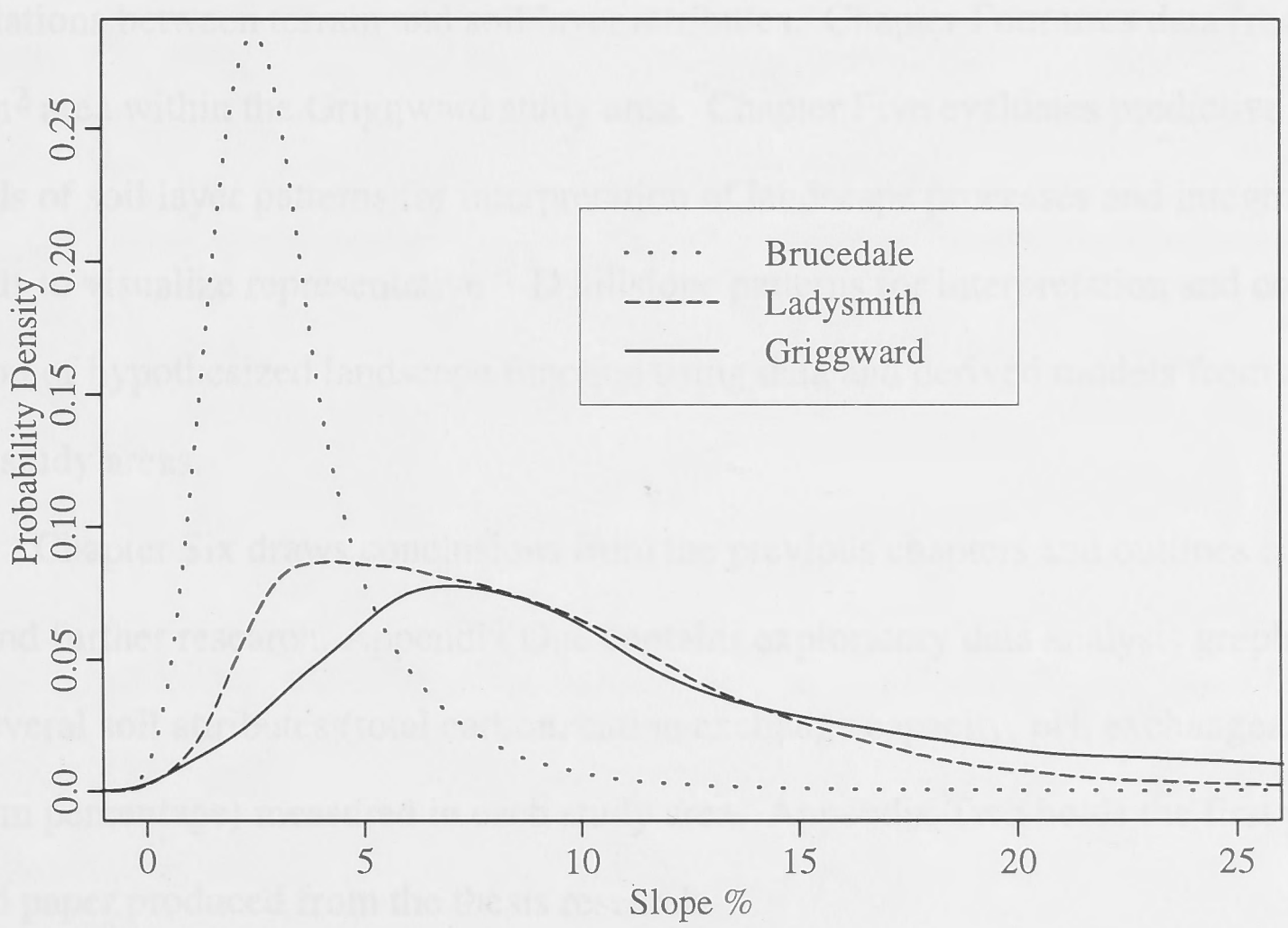


Figure 1.2 Study Area Slope Attribute Distributions

1.3 THESIS ORGANIZATION

In development of the basic aim stated above, the thesis chapters are ordered to build the ideas, introduce the tools and demonstrate their application with data from three study areas. Chapter Two reviews pertinent literature to refine ideas and concepts, introduce tools and submit a framework for implementation. Individual chapter hypotheses are then proposed and tested in Chapters Three, Four and Five. These chapters are individually sectioned as:

- introduction;
- material and methods;
- results and discussion;
- conclusions; and
- references cited.

Chapter Three delves into the mechanics of the sampling, exploratory data analysis (EDA) and statistical modelling process for spatial prediction and visualization of individual models using data from the Ladysmith study area. Chapter Four is an empirical study of the influence of DEM resolution on terrain attribute distributions and correlations between terrain and soil layer attributes. Chapter Four uses data from a 16 km^2 area within the Griggward study area. Chapter Five evaluates predictive models of soil layer patterns for interpretation of landscape processes and integrates models to visualize representative 3-D hillslope patterns for interpretation and comparison of hypothesized landscape function using data and derived models from all three study areas.

Chapter Six draws conclusions from the previous chapters and outlines continued and further research. Appendix One contains exploratory data analysis graphics for several soil attributes (total carbon, cation exchange capacity, pH, exchangeable sodium percentage) measured in each study area. Appendix Two holds the first published paper produced from the thesis research.

An assumption of this thesis is that the reader is familiar with basic concepts of pedology, statistics and geographical information systems.

1.4 REFERENCES CITED

Ahnert, F. 1988. Modelling landform change. p. 375-400. *In* M.G. Anderson (ed.) Modelling geomorphological systems. Catena Suppl. 6, Catena Verlag Publ., Cremingen-Distadan, the Netherlands.

Austin, M.P. and N.J. McKenzie. 1988. Data analysis. pp.210-231. *In* R.H. Gunn, J.A. Beattie, R.E. Reid and R.H.M. van de Graaff (eds.) Australian Soil and Land Survey Handbook. Inkata, Melbourne.

Bell, J.C., R.L. Cunningham, and M.W. Havens. 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. Soil Sci. Soc. Am. J. 56:1860-1866.

Bell, J.C., R.L. Cunningham, and M.W. Havens. 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Sci. Soc. Am. J. 58:464-470.

Bouma, J. 1989. Land qualities in space and time. p.3-13 In J. Bouma and A.K. Bregt (eds.) Land Qualities in Space and Time. Pudoc, Wageningen, The Netherlands.

Butler, B.E. 1964. Can pedology be rationalized? A review of the general study of soils. Australian Soc. Soil Sci. Publication No. 3. Canberra, Australia.

Butler, B.E. 1980. Soil classification for soil survey. Clarendon Press, Oxford.

Burrough, P.A. 1986. Principles of geographical information systems. Clarendon Press, Oxford.

Chambers, J.M. and T.J. Hastie. 1992. Statistical models in S. Wadsworth & Brooks, Los Angeles.

Christian, C.S., and G.A. Stewart. 1968. Methodology of integrated surveys. p. 233-280. *In* Aerial Surveys and Integrated Studies. Proc. Toulouse Conf. 1964. UNESCO.

Dent, D., and A. Young. 1981. Soil survey and land evaluation. Allen & Unwin, London.

Dikau, R. 1989. The application of a digital relief model to landform analysis in geomorphology. p51-78. *In* J.D. Raper (ed.) Three dimensional applications in geographical information systems. Taylor & Francis, London.

Dobrovol'skiy, G.V. 1994. The development of studies on the structure of the soil cover as a branch of soil geography. Eurasian Soil Sci. 26(4):1-9.

Hewitt, A.E. 1993. Predictive modelling in soil survey. Soils Fert. 3:305-314.

Hole, F.D., and J.B. Campbell. 1985. Soil landscape analysis. Rowman & Allenheld,

Totowa.

Holmgren, G.G.S. 1988. The point representation of soil. Soil Sci. Soc. Am. J. 52:712-716.

Hudson, B. 1992. The soil survey as a paradigm-based science. Soil Sci. Soc. Am. J. 56:836:841.

Huggett, R.J. 1975. Soil landscape systems: a model of soil genesis. Geoderma. 13:1-22.

Jenny, H. 1941. Factors of soil formation: a system of quantitative pedology. McGraw-Hill, New York.

Mabbutt, J.A. 1968. Review of concepts of land classification. p. 11-29. *In* G.A. Stewart (ed.) Land Evaluation. Macmillan, Melbourne.

McKenzie, N.J., and M. Austin. 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329-355.

McSweeney, K., P.E. Gessler, B. Slater, R.D. Hammer, J. Bell, and G.W. Petersen. 1994. Towards a new framework for modelling the soil-landscape continuum. p.127-145. *In* Factors of soil formation: a fiftieth anniversary retrospective. SSSA Special Pub. 33. Madison, WI.

Moore, I.D., P.E. Gessler, G.A. Neilsen, and G.A. Peterson. 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57:443-452.

Northcote, K.H. 1984. Soil-landscapes, taxonomic units and soil profiles: a personal perspective on some unresolved problems in soil survey. Soil Surv. Land Eval. 4:1-7.

Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63:197-214.

Pawluck, R.R., J.A. Sandor, and J. Tabor. 1992. The role of indigenous soil knowledge in agricultural development. J. Soil and Water Cons. 47:298-302.

Shovic, H.R. and Montagne, C. 1985. Application of a statistical soil-landscape model to an Order III wildland soil survey. Soil Sci. Soc. Am. J. 49:961-968.

Slater, B.K., K. McSweeney, S.J. Ventura, B.J. Irvin, and A.B. McBratney. 1994. A spatial framework for integrating soil-landscape and pedogenic models. p. 169-185.C *In* Quantitative modeling of soil forming processes. SSSA Special Publication 39. Madison, WI.

Soil Survey Division Staff. 1993. Soil survey manual. Agriculture Handbook No. 18. U.S.D.A. U.S. Government Printing Office. Washington, D.C.

Thompson, C., and A.W. Moore. 1984. Studies in landscape dynamics in the Cooloola-Noosa River area, Queensland. I. Introduction, general description and research approach. CSIRO Aust. Div. Soils. Divisional Report No. 73. Adelaide, Australia.

Zimmerer, K.S. 1994. Local soil knowledge: answering basic questions in highland Bolivia. J. Soil Water Cons. 49:29-34.

Chapter Two: Literature Review & Concept Development

'Truth comes out of error more easily than out of confusion'

Francis Bacon (1561-1626)

2.1 BROAD PRINCIPLES

In a review of the general study of soils, Butler (1964) stated: "Because of the lack of reality about soil entities, it is necessary that general studies in soils should be carried on in terms of soil attributes, not of soil types, or even of soil profiles." Beckett and Webster (1971) showed that many soil attributes exhibit different scales of variation and that soil survey mapping units, based on soil morphological attributes, often capture only a small percentage of this variation for any particular soil attribute. McSweeney *et al.* (1994) suggest that soil complexity has been rationalized using subjective and generalized taxonomic concepts and mapping procedures because of the lack of computer-based tools for organizing and analysing data. Many have concluded that in the course of general purpose soil survey, invaluable field data are collected and derived mental models are developed, but lost or obscured in published product because of the rigidity imposed by cartographic conventions and scale of map publication (Burrough, 1986; Bell *et al.* 1992; Hewitt, 1993; McSweeney *et al.* 1994). Geostatistics provides methods for quantifying individual soil attribute spatial variation and methods for interpolation of sample data (Webster and Oliver, 1990). However, the methods have not been widely adapted for use in general soil survey perhaps because of the complexity of the techniques, requirements for large amounts of data at appropriate lag spacings and difficulty of incorporating useful environmental correlations and landscape process understanding.

The intention of this review is not to rehash the many works that have critiqued conventional methods (Gibbons, 1961; Beckett and Bie, 1978; Butler, 1980;

Webster and Oliver, 1990; McKenzie, 1991; Bell *et al.* 1992; Moore *et al.* 1993; Hewitt, 1993; McSweeney *et al.* 1994) but to focus on three main criticisms of conventional survey. These are:

- conventional methods are usually not explicitly stated (e.g. measurement scales, sampling strategy, data analysis);
- derived soil-landscape models are not quantitatively expressed (e.g. soil map units with no quantification of relationships used or uncertainty and error for predicting particular attributes); and
- the product is difficult to interpret and use by non soil scientists (e.g. map units based on a specialized taxonomy rather than predictions of individual attributes).

Hewitt (1993) argues that predictive modelling in soil survey must meet these criteria if it is to be regarded as scientific.

This review is selective and does not cover all potential conceptual implementations or new tools that may be used for soil-landscape modelling. The specific focus is on integration of concepts and techniques useful for exploring and utilizing a broad range of environmental correlations for developing and extrapolating statistical soil-landscape models for spatial prediction. Implementation of an explicit and quantitative modelling approach requires a systematic documentation and answering of the following questions.

- At what scale(s) are primary environmental variables available?
- What is the intended scale of soil-landscape model application?
- How is the broader soil population stratified to define the model scope?
- What are the criteria for selection of field sample locations?
- What tools may be used to explore environmental correlations?
- How do we confirm, quantify and evaluate these correlations? and
- How do we implement models for spatial prediction and extension to users?

This review addresses each question and uses the following sequence:

- scale, stratification and model scope;

- field sampling;
- exploratory data analysis;
- statistical modelling; and
- spatial prediction.

2.2 SCALE, STRATIFICATION, AND MODEL SCOPE

An important initial task in the development of statistical soil-landscape models is stratification of the continuous soil population to define the spatial and conceptual domain within which a particular model applies (McSweeney *et al.* 1994). McCullagh and Nelder (1989) refer to this as the "model scope". Conventional soil survey often uses an implicit state factor approach (i.e. Jenny, 1941; 1980) where physiographic areas and mapping units are defined using a qualitative stratification of environmental variables (e.g. climate, geology, landform, airphoto pattern, classified soil profile).

McSweeney *et al.* (1994) laid out a conceptual framework for developing soil-landscape models by descending levels of scale from the broad region or physiographic domain to the local scale where measurements of specific soil attributes are made. This suggests cascading levels of stratification towards the scale of application that progressively encompasses and narrows variation for the attribute of interest using different environmental variables. For example, the broad physiographic domain may be defined by bedrock geology type from an upper meso-scale geology map (e.g. 1:100 000 cartographic scale). The next scale stratification may be at the mid meso-scale catchment and hillslope level using landform as quantified by digital terrain attributes (20m grid spacing) within a bedrock geology type. The next level may be a soil layer or horizon over hillslopes measured from a sampling of soil pits or soil cores followed by the basic measurements of the variable of interest (e.g. pH, total Carbon, cation exchange capacity) within the soil horizon. In concept, this approach

enables integration of a variety of different environmental stratifiers, attribute types and measurement scales (e.g. geology = nominal, landform = ratio scale, soil layer = ordinal, pH = interval) whose explicit use in a statistical model depends on the variation accounted for in the response sample set for the variable of interest.

An important assumption of this approach is that the sample set for the variable of interest is representative of the variation within the spatial area encompassed by the stratifier variable. If this assumption is valid, the model scope is defined by the ranges of the predictor environmental variables used.

Complementing the concepts of McSweeney *et al.* (1994), Hoosbeek and Bryant (1992), drawing on the work of Dijkerman (1974), propose a hierarchy of soil systems ranging in levels of complexity and organization from the soil region to molecular interaction with the soil pedon the central *i* level. In discussion of scale and measurement for development of quantitative models of pattern and structure in ecological systems, Allen and Hoekstra (1992) suggest that: "For an adequate understanding leading to robust prediction , it is necessary to consider at least three levels at once: 1) the level in question; 2) the level below that gives mechanisms; and 3) the level above that gives context, role or significance."

To synthesize these concepts, it is useful to define a theoretical soil-landscape model with defined components to provide a framework for the rest of the thesis. This may be considered a re-formulation of Jenny's state factor equation (Jenny, 1941). The intended level of application for soil-landscape modelling in this thesis is the lower meso-scales useful for land management. The theoretical model is defined as:

$$\mathbf{S}i_{(1...n)} = fi_{(1...n)}\,(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots \mathbf{x}_n) \qquad (2.1)$$

where:

S is a response soil attribute (e.g. A horizon depth, profile total carbon, B horizon clay %)

i_n represents the explicit definition of the model domain or scope using available environmental variables (e.g. a region defined using

climate, geology, landform)

f is the statistical model (e.g. generalized linear, generalized additive, tree-based, geostatistical models)

$\mathbf{x_n}$ is a predictor environmental variable (e.g. slope, catchment position, drainage area, solar radiation)

The i_n and $\mathbf{x_n}$ environmental variables are listed here as separate components due to differences in supporting data and scale. The i_n variables are available at the upper meso- to macro- scales and are used for broad environmental stratification, the level above that gives context for soil survey. The $\mathbf{x_n}$ variables are available at the lower meso- and local hillslope scales or the level of intended application. For purposes of spatial prediction, it is important that the $\mathbf{x_n}$ variables be available in a spatially continuous manner. Statistical models f are often defined using correlations between the $\mathbf{x_n}$ variables and measurements of soil variables, $\mathbf{s}$, made in the field or lab. The $\mathbf{s}$ soil variables are usually obtained through field sampling, are usually not available in a spatially continuous manner, and are at a scale or support level below the intended application. These data can often be related to mechanisms and soil-landscape processes.

For example, climatic and geologic (i_n) variables are often available at broad regional scales (e.g. 1:100 000 cartographic scale). However the spatial distribution and dynamics of water, temperature and weathered rock or biogeochemical material are locally modified by landform. Although it is difficult and expensive to measure climatic and geologic variables at local scales, easily measured digital terrain variables ($\mathbf{x_n}$) may be used as surrogates to sample, explore and integrate local effects caused by landform into a statistical model f for predicting a soil variable $\mathbf{s}$. Exploration of environmental correlations relies on samples of the soil variable ($\mathbf{s}$) often collected at micro-scales via described soil layers, depth increments or using a bulking strategy.

Ideally, measurements of a range of environmental and soil variables would be available for a range of scales across a region, but this is never logistically feasible. Hence, environmental stratification, environmental correlation and definition of model scope may be performed in many different ways depending on survey purpose, physiographic characteristics of the area and the availability of environmental data. An advance would be to explicitly state the scale level of all environmental variables used for stratification, prediction and definition of the model domain. The scope and amount of variation accounted for in a soil-landscape model depend on the explicit relationships defined by equation 2.1 using sample evidence.

2.3 FIELD SAMPLING STRATEGY

If we assume stratification of the broader soil population can be made using explicitly defined i_n environmental variables (e.g. bedrock geology map units), explicit decisions are then required on how to sample within the spatial domain for exploration of environmental correlations and development of soil-landscape models. Because a central application of a soil-landscape model is spatial prediction, it is also important that samples are spread or optimized in geographic space to capture variation and provide a reasonable representation of the spatial domain. This also enables evaluation of the appropriateness of the i_n environmental stratification.

In conventional or free soil survey (Steur, 1961; Beckett, 1968), ground observations are irregularly located according to the surveyors judgement (Reid, 1988). Without documentation of the surveyors decisions it is impossible to test or repeat schemes. O'Brien (1992) states: " The quality of inference depends on how adequately the sample represents the population. If the sample is some sort of microcosm, a population in miniature, inference is likely to be reasonably accurate. If, however, the sample is wholly arbitrary, or has been gathered without respect for known features of the population, inference is likely to be of limited value." Allen

and Hoekstra (1992) suggest that scales of perception may exist where phenomena become simpler and predictions are improved if the characteristics of the material system are used to anchor the investigation. For example, if the hillslope is the functional unit for re-distribution of material and energy within a particular soil-landscape system, it is sensible to attempt incorporation of environmental variables that capture variation at this meso-scale and perhaps use these variables to guide sampling. Within the context of general purpose soil survey, these concepts suggest that we may develop more robust models by integrating pedologic and landscape process understanding with traditional and spatial statistical theory to develop an explicit sampling strategy.

2.3.1 Environmental Gradients

Use of environmental gradients to guide sampling has existed in the soil and plant ecology fields for a long time (Jenny, 1941; Curtis, 1959; Vitousek, 1994). In review of his final chapter, Jenny (1980) states: "pedogenic order in a landscape is unraveled by stratified random sampling along vectors of the state factors." Gillison and Brewer (1985) used gradsect sampling or the deliberate selection of transects which contain the steepest environmental gradients in an area to ensure capture of the full range of variation in vegetation (Austin and Heyligers, 1989).

Moore *et al.* (1993) discovered many useful relationships between quantitative terrain attributes and soil patterns and suggested that terrain attributes should be used as a guide for soil survey sampling in un-mapped territory. This blends with the concept of a provisional predictive pedologic model (McKenzie and Austin, 1993) where an *a priori* model is hypothesized for empirical testing, improvement and discovery of other useful environmental correlations. This is similar to the process a traditional field surveyor would use, but is placed on an explicit and quantitative foundation.

Environmental variables ($\mathbf{X_n}$) may be used as surrogates for the model domain soil population and sampling along gradients in environmental attribute space may assist in capturing the full range of soil variation, maximize useful environmental correlations and hence, the quality of spatial prediction.

2.3.2 Terrain

Terrain or landform has long been recognized as an important modifier of material and energy fluxes that influence soil formation (Milne, 1935; Jenny, 1941; Watson, 1965; Hole and Campbell, 1985) and is widely used in descriptive soil-landscape models throughout the literature (Ruhe, 1975; Conacher and Dalrymple, 1977). It is usually the most useful lower meso-scale predictor of soil variation in surveys.

Several workers have recently called for explicit incorporation of quantitative terrain attributes that describe water movement, catchment hydrology and landscape processes into the soil-landscape modelling process (Moore *et al.* 1991; Hall and Olson, 1991; Daniels and Hammer, 1992; Moore *et al.* 1993; McSweeney *et al.* 1994). The use of quantitative terrain attributes for soil-landscape modelling is not new (Troeh, 1964; Walker *et al.* 1968; Speight, 1968, 1974; Odeh *et al.* 1990; 1994a; Moore *et al.* 1993; McKenzie and Austin, 1993: Bell *et al.* 1992, 1994). However, none of these papers integrate terrain variables with traditional and spatial statistical theory to design a sampling strategy that facilitates exploration of environmental correlations for quantitative soil-landscape modelling.

Moore *et al.* (1991) review the computation of terrain attributes from digital elevation models and discuss applications in the earth sciences. They describe primary and secondary or compound terrain attributes, drawing on earlier work of Speight (1968, 1974) and Evans (1971) in development of quantitative geomorphometry. Table 2.1 lists a range of terrain attributes that may be computed from a DEM. Primary terrain attributes are those that can be directly calculated from a DEM and include the first and second derivatives (slope, aspect, plan curvature, profile

Table 2.1 Digital Terrain Attributes (from Speight, 1974; Moore *et al.*, 1991)

Attribute	Definition	Significance
Primary Terrain Attributes		
Altitude	Elevation	Climate, vegetation, potential energy
Slope	Gradient	Overland and subsurface flow velocity and runoff rate, precipitation, vegetation, geomorphology, soil water content, land capability class
Aspect	Slope azimuth	Solar insolation, evapotranspiration, flora and fauna distribution and abundance
Profile curvature	Slope profile curvature	Flow acceleration, erosion/deposition rate, geomorphology
Plan curvature	Contour curvature	Converging, diverging flow, soil water content, soil characteristics
Secondary Terrain Attributes		
Flow path length	Maximum distance of water to a point in the catchment	Erosion rates, sediment yield, time of concentration
Catchment area	Area draining to catchment outlet	Runoff volume
Specific catchment area	Upslope area per unit width of contour	Runoff volume, steady-state runoff rate, soil characteristics, soil water content,
Upslope height	Mean height of upslope area	Potential energy
Upslope slope	Mean slope of upslope area	Runoff velocity
Dispersal slope	Mean slope of dispersal area	Rate of soil drainage
Catchment slope	Average slope over catchment	Time of concentration
Upslope area	Catchment area above a length of contour	Runoff volume, steady-state runoff rate
Dispersal area	Area downslope from a short length of contour	Soil drainage rate, geomorphology
Upslope length	Mean length of flow paths to a point in the catchment	Flow acceleration, erosion rates
Dispersal length	Distance from a point in the catchment to the outlet	Impedance of soil drainage
Catchment length	Distance from highest point to outlet	Overland flow attenuation
Wetness index	Computed using specific catchment area and slope	Soil moisture
Streampower index	Computed using specific catchment area and slope	Erosive power of overland flow
Sediment transport capacity index	Computed using specific catchment area and slope	Erosion and deposition processes
Solar radiation indices	Computed using terrain, climate surface reflectance data	Energy availability and flux

curvature) as well as flow direction. Secondary or compound terrain attributes are derived from combinations of the primary attributes and can often be related to landscape processes or hydrological and catchment context (Bevan and Kirkby, 1979; O'Loughlin, 1986; Moore *et al.*, 1991). Digital terrain attributes quantify the

geometry of the landsurface and provide a quantitative analogue to the more qualitative process of airphoto interpretation used routinely in soil resource inventory. Moore *et al.* (1993) found several useful correlations between digital terrain and sampled soil attributes and used regression models to develop spatial predictions. The most useful terrain attribute was the wetness index. Moore *et al.* (1994) renamed this secondary terrain attribute the compound topographic index (Moore *et al.* 1994; Gessler *et al.* 1995) because of the soil transmissivity assumptions required in use of the term "wetness" as originally developed in the hydrology literature (Bevan and Kirkby, 1979).

McSweeney *et al.* (1994) suggest that once initial broad scale environmental stratification of the soil population is completed, terrain or digital elevation models (DEM's) may be used as the basis for spatially organizing data with reference to the three-dimensional ground surface to develop locally explicit soil-landscape models. McSweeney *et al.* (1994) recognized that external landform is not always a good predictor of soil patterns because of parent material variations, bedrock structure, water table expression, stratigraphic breaks, relic and catastrophic features and others (Hole and Campbell, 1985; Gerrard, 1990; Daniels and Hammer, 1992; McKenzie and Austin, 1993). The key point is that external landform is often an inexpensive first approximation or predictor of soil spatial patterns useful for provisional model development and landscape sampling. With sufficient sampling numbers, data exploration should uncover inconsistencies or changes in predictive relationships over external landform that may require further investigation and limit the cost-effective use of landform.

2.3.3 Soil Layers

Soil layers are defined by field sampling and descriptive morphology and may have a pedogenic (soil horizon) or geomorphic (stratigraphic unit) origin. They are the key field indicator for changes in soil patterns. Description of the horizon sequence or soil profile has been the basis for rationalizing soil complexity for

taxonomic classification and map unit delineation in conventional survey. This assumes good correlation between qualitative morphological attributes and a range of other soil attributes (e.g. chemical, physical, biological). However, quantitative and explicit supporting evidence for this assumption has been lacking.

The importance and usefulness of soil layers or horizons as predictors of soil variation and soil attribute patterns is well documented (Fitzpatrick, 1967; Butler, 1982; Bouma, 1989; Fitzpatrick, 1993; McKenzie and Austin, 1993; McSweeney *et al.* 1994; van Wesenbeeck and Kachanowski, 1994; Slater, 1994; Price, 1994; Gessler *et al.* 1995). McBratney (1993) calls for a new paradigm for soil modelling based on the soil horizon as the fundamental spatial entity in contrast to taxonomic classification approaches that lump horizons effectively discarding potentially useful lower meso-scale predictors. Use of soil layers has merit because they are simple and inexpensive to describe in the field and are often an efficient classifier of complex multivariate data. Furthermore, the soil layers at any location are a result of integrated pedo-geomorphic and hydrological processes (Simonson 1959, Butler, 1964; Jenny, 1980). As such, a description of the arrangement, dimension and nature of the soil layers at locations in the landscape may be used as a link or pointer to the spatial distribution of landscape processes important for interpreting soil attribute patterns.

Soil layers also provide a logical building block for spatial modelling, visualization and interpretation of how sequences of layers behave in the landscape (McKenzie and Austin, 1993; Moore *et al.* 1993; Slater *et al.* 1994; McSweeney *et al.* 1994). McSweeney *et al.* (1994) suggest that soil layers are useful spatial entities that may be aggregated, where appropriate, for modelling different soil attributes in the landscape. The key point is that soil layers may be useful and should be preserved for evaluation as a potential lower meso-scale or upper micro-scale stratifier and predictor of soil variation. However, qualitative decisions on the definition and identification of layers in the field remains an issue and more explicit and repeatable procedures are required.

2.3.4 Field Data Collection and Attribute Measurement

Measurement of individual soil attributes in the field or laboratory are the response variables (**s**) on which environmental correlations and soil-landscape models are based. They are usually described or measured in the field or laboratory on micro-scale samples collected from a limited number of locations. Accurate locationing of field collection sites is key to obtaining values of other potential predictor environmental variables over the same geographic space. Due to many logistical constraints (e.g field conditions, field apparatus, laboratory apparatus, cost of analysis, time required) field data collection is often the most expensive part of general purpose soil survey (Bie and Beckett, 1970; Dent and Young, 1981). Reid (1989) reports that field work normally accounts for 60-70% of the total cost of a survey.

Measurement itself also imposes constraints. Allen and Hoekstra (1992) closely couple the issue of scale and perception of variation in environmental variables with observation measurement indicating there is always an imposition of an attribute scale when quantitative measurements are made. Bouma (1989) has emphasized similar ideas in the soils literature by highlighting the importance of sampling volume in the measurement of soil attributes such as hydraulic conductivity (Bouma, 1983; Lauren *et al.* 1988). This is particularly relevant for soil attribute measurements because sampling often disturbs the natural setting of the soil-landscape continuum. Webster and Oliver (1990), drawing on geostatistical research (Journel and Huijbregts, 1978), define the term sample *support* as "the dimensions of the individual - its size, shape and orientation" and suggest that it should always be stated for soil variables when reporting results.

Thus the cost of field data collection and measurement substantiates the importance of careful selection of field sample locations to develop environmental correlations with spatially continuous attributes that are simple to measure. Likewise, the specific techniques and imposition of measurement scales will influence how we

quantify variation, correlate variation between different environmental variables and report model confidence and uncertainty based on residuals or prediction error using sample evidence. This substantiates the use of explicit statistical methods that quantify observed relationships.

2.3.5 Statistical Sampling

As stated above, the central application of a soil-landscape model in this thesis is spatial prediction. General purpose soil survey often seeks to gather data for a range of soil attributes from a single sampling without *a priori* knowledge of specific soil attribute variance. Different soil attributes exhibit disparate scales of univariate (e.g. attribute space) and spatial (e.g. geographic space) variation which complicates the development of probabilistic sampling approaches (e.g. geostatistical) that require prior knowledge of variance. The approach developed here suggests that a provisional model (McKenzie and Austin, 1993) be based on an environmental gradient(s) (e.g. meso-scale terrain) to define the attribute space and spread samples in both attribute and geographic space.

To explicitly and quantitatively use environmental gradients requires characterization of the chosen environmental attribute and geographic spaces defined by a study or survey area. Figure 1.2 showed univariate probability density functions (pdf's) that quantify the range, shape and variation of the slope environmental attribute spaces for the three study areas. Parameters that characterize the pdf's, such as statistical moments (e.g. mean, median, variance, skewness, kurtosis) and quantile measures, can be used to explicitly segment attribute space (Snedecor and Cochran, 1980) . Randomization is also important to ensure that each member of the population has an equal chance of being selected (Snedecor and Cochran, 1980; Webster and Oliver, 1990). Together, this results in a stratified random sampling in environmental attribute space analogous to Jenny's sampling along vectors of state factors. Box *et al.* (1978) discuss the use of response surface methods that may be

useful if the intention is to sample across several environmental gradients simultaneously.

A complement to the pdf that quantifies scale of spatial variation for an environmental variable is the variogram (Webster and Oliver, 1990; Cressie, 1991). The variogram is a graphical tool of spatial statistics that indicates the range or distance within which spatial dependence occurs. Samples spread beyond the range distance do not exhibit spatial correlation in geographic space and therefore maximize the spatial representation and potential information gained by each sample location (Webster and Oliver, 1990). Integrating concepts about how to allocate sample locations in both environmental attribute and geographic space provides initial explicit guidance for sampling density and total number of samples required over a study area. This is equivalent to Allen and Hoekstra's (1992) concept of using characteristics of the material system to anchor the investigation.

For example, a slope pdf, as defined in Figure 1.2 can be segmented according to five evenly spaced quantiles along the gradient (e.g. 0-20%, 20-40%, 40-60%, 60-80%, 80-100%). A variogram computed from these same data may show that spatial independence occurs beyond 500 metres. This suggests that an individual sample in slope attribute space is representative of an area with a 500 metre radius. Therefore, samples that are close in slope attribute space (e.g. the 0-20% quantile samples) should be spaced a minimum of 500m apart in geographic space or use an exclusion circle with a radius of 500m or area of about 78 hectares. If ten samples are planned for the 0-20% quantile, the study area should be no smaller than 780 hectares and preferably much larger to enable random selection of 0-20% quantile patches over a broader region defined by the upper level stratification.

An assumption of such an integrated spatial and environmental gradient sampling approach is that an $\mathbf{x_n}$ environmental variable(s) can be selected to usefully match and stratify the variation for a range of soil variables ($\mathbf{s}$). Exploration of

collected field sample data will suggest whether or not $\mathbf{X_n}$ and $\mathbf{S}$ variables exhibit useful correlations. If the survey also requires information about short range variation for a particular soil variable, it may be appropriate to subsequently design a specific sampling at nested scales either within the range of spatial dependence or as a separate sampling if environmental correlations do not exist (Webster and Oliver, 1990).

Although the approach here has not been previously tested or reported in the literature, it is explict and quantitative, placing it on scientific grounds for testing, evaluation and improvement.

2.4 EXPLORATORY DATA ANALYSIS

Many possible environmental predictors and combinations are possible for developing statistical soil-landscape models. The task of comprehensively evaluating and identifying them is non-trivial. Therefore exploratory data analysis (EDA) is an essential first step and critical tool for evaluation of provisional models and identification of other important correlations.

EDA (Tukey, 1977) involves the graphical exploration of data to detect outliers, trends or groups and evaluate statistical assumptions (Austin and McKenzie, 1988; Cleveland, 1993). Austin and McKenzie (1988) indicate that one of the main goals of exploratory data analysis is for hypothesis generation to guide subsequent confirmatory data analysis or statistical modelling. Cleveland (1993) discusses data visualization as a different and complemetary paradigm to the traditional foundations of probabilistic inference laid out by Fisher (1958). Cleveland (1993) suggests that with a knowledge of the subject under study, the two components, graphing and fitting, can sometimes replace the need for probabilistic inference. He further states: "In other cases, visualization is not enough and probabilistic inference is needed to help calibrate the uncertainty of a less certain issue. When this is so, visualization has

yet another role to play - checking assumptions." Cleveland (1993) discusses data visualization tools according to data type (univariate, bivariate, trivariate, hypervariate and multiway). A brief summary of new methods of EDA useful for the analysis of soil survey data are discussed below. Use of these methods will be demonstrated in subsequent chapters.

2.4.1 Univariate EDA

Univariate data are measurements of a single quantitative variable (Cleveland, 1993). Assumptions required of classical methods of statistical inference are an outlier-free and nearly normal distribution with uncorrelated observations. Tukey (1977) introduced the stem and leaf plot as a simple way to summarize and display information about a univariate sample.

A more powerful way to visualize and compare sample distributions and identify outliers is through the use of quantiles (Wilk and Gnanadesikan, 1968). Cleveland (1993) states: "The f quantile, q(f), of a set of data is a value along the measurement scale of the data with the property that approximately a fraction f of the data are less than or equal to q(f)." Distributions can be compared by viewing quantiles of the same f value from each distribution un-influenced by differences in the actual number of samples. Hence, any sample dataset or transformations of a dataset can be graphically compared to the normal distribution, a range of other known probability distributions or distributions of other sample datasets. The quantile-quantile plot or Q-Q plot is the graphical method for doing this and is discussed in detail by Cleveland (1993). The boxplot (Tukey, 1977) is another way of distilling the information contained in a Q-Q plot. These methods complement traditional sample distribution display methods such as frequency histograms, probability density and cumulative frequency functions as well as numerical measures of statistical moments (e.g. mean, standard deviation, skewness, kurtosis). These techniques may also be used to examine residuals from fitted models.

2.4.2 Bivariate EDA

Bivariate data are paired measurements of two quantitative variables (Cleveland, 1993). The classical method of bivariate EDA is the scatterplot. Cleveland (1993) suggests the addition of smooth curves to the scatterplot to enhance the perception of the pattern of dependence. He discusses the importance of aspect ratio or data banking for improving display, particularly when evaluating complex datasets that include many sample displays. These methods may also be used to evaluate model residuals.

Pairwise scatterplot matrices (Cleveland, 1993) may be used for visual comparison of bivariate relationships for a broad range of variables at once. Moore *et al.* (1993) presented a pairwise scatterplot-correlation matrix summarizing the relationships between soil and terrain variables through scatterplots in panels below the diagonal and correlation coefficients in panels above the diagonal. Boxplots may also be used in a bivariate and multivariate manner to visualize relationships between nominal or ordinal attributes and continuous variables.

2.4.3 Multivariate EDA

Multivariate EDA methods are used for displaying and exploring measurements of three or more quantitative variables. A large number of multivariate techniques exist and many have been used to explore soil data (e.g. similarity measures, ordination, discriminant analysis, hierarchical and non-hierarchical numerical classification) without gaining wide acceptance or use (Webster and Oliver, 1990). Therefore discussion here will focus on newer methods that show promise for exploration of environmental relationships as introduced by Cleveland (1993) .

"Conditioning" (Cleveland, 1993) is a powerful technique for studying how a response variable depends on two or more factors or variables and forms the basis for a number of graphical methods (Yates, 1937; Snee, 1985; Tukey and Tukey, 1981). A conditioning plot or coplot (Cleveland, 1993) presents conditional dependence in a visually efficient way by integrating many of the previously discussed methods into a

single display enabling exploration and discovery of complex variable interactions. This allows the presentation of a matrix of scatter plots (a vs. b variables) with the x and y axes of the overall matrix organized or conditioned along the scale of one or two additional variables (c & d variables). Smooth fit lines, data banking and variable scale grids may be added to improve the perception of variations in attribute space (Cleveland, 1993). Another class of useful methods discussed by Cleveland (1993) is direct manipulation graphics (Cleveland and McGill, 1988) such as brushing (Becker and Cleveland, 1987) and spinning (Fisherkeller *et al.* 1988). Brushing allows the interactive and dynamic highlighting of data points or groups throughout an entire matrix of scatterplots to explore multivariate relationships. Spinning enables the interactive movement of axes for viewing the three dimensional nature of attribute space of different variable combinations.

Tree-based models are an exploratory technique for uncovering structure in data (Clark and Pregibon, 1992) based on the methods outlined by Breiman *et al.* (1984). Although not considered amongst the EDA techniques discussed by Tukey (1977) or Cleveland (1993), they provide a complementary way of exploring multivariate data that do not suffer from the same parametric assumptions of other techniques. Tree-based models will be discussed in more detail in the next section.

Exploratory data analysis is used throughout the process of statistical soil-landscape model development and refinement. EDA is used to:

- initially check whether normal univariate assumptions are valid;
- explore structure and relationships in data to guide modelling efforts; and
- check the validity of model assumptions or appropriateness through the visualization of residuals and other diagnostics.

Austin and McKenzie (1989) distinguish exploratory data analysis from confirmatory data analysis or statistical modelling. While discussed here as separate processes, both are used in an integrated and complementary manner for soil-landscape modelling.

2.5 STATISTICAL MODELLING

2.5.1 Background

Scientific models expressed mathematically are central to studying natural phenomena (Chambers and Hastie, 1992). Statistical modelling is one approach for developing useful mathematical expressions that describe pattern in data and enable definition of environmental correlations for predictive soil-landscape modelling. Many different approaches to statistical modelling exist for predictive modelling with soil data (Webster and Oliver, 1990; O'Brien, 1992; Chambers and Hastie, 1992). Several recent works have demonstrated the use of numerical classification techniques (Bell, 1990; Odeh, 1990; Slater, 1994). The aim here is to introduce and discuss other complementary methods with a focus on techniques that have not been widely reported in the soils literature and enable simple construction of environmental correlations for more extensive spatial prediction.

Broadly stated, parametric methods assume a known underlying probability process (usually normal) while non-parametric methods make no such assumption and may be used when the data distribution is far from normal (O'Brien, 1992; Statistical Sciences, 1993). If parametric methods are appropriate, this opens up a broad range of classical statistical methods and theory based on the underlying probability process (O'Brien, 1992). However, modern statistical methods include a broad array of techniques that expand traditional methods, span the border between parametric and non-parametric methods or use alternative approaches. The goal in development of a statistical model, f, is to explain as much variation in the response, $\mathbf{s}$, as possible while attempting to keep the model simple by minimizing the number of parameters requiring estimation (Chambers and Hastie, 1992). If environmental correlations or knowledge of landscape processes are useful, they may be incorporated into the sampling and model building process. EDA techniques of graphing and fitting

(Cleveland, 1993) may point to short-cuts for modelling. The appropriate method will depend on the type of response variable, $\mathbf{s}$ (e.g. continuous, nominal, binary) and identification of useful relationships with measured $\mathbf{x_n}$ predictor variables. This section provides an overview and simple comparison of a progression of statistical techniques that may be used to define f (2.1).

2.5.2 Linear Models

The statistical use of linear models goes back to Laplace and Gauss in the early nineteenth century and continues to underlie much of statistical modelling (Stigler, 1986). The principal technique is that of linear least-squares regression which is an objective and efficient method of determining the "best-fit" straight line (Stigler, 1981). This technique estimates the parameters of a straight line fit by minimizing the squared sum of residuals. The assumptions of these methods are that the underlying relation is linear and the residuals are independent and normally distributed with constant variance. Appropriateness of a fit is evaluated by reduction in variance and analysis of model residuals. Residual patterns provide the basis for improving the fit (Cook and Weisberg, 1982). Transformations of the response or predictor variables may be performed to improve linear relationships and avoid assumption violations. Multiple regression refers to a linear model of relationships where the response depends on two or more predictor variables (Weisberg, 1980; Draper and Smith, 1981). Robust regression methods (Statistical Sciences, 1993) are an extension of the linear model that perform equivalently to linear models for normally distributed data, but incorporate weighting criteria when the errors are not normally distributed and outliers or high leverage points exist (Cook and Weisberg, 1982). Polynomial regression is a method where higher orders of a predictor variable may play the role of a single predictor in the process of least squares regression.

2.5.3 Generalized Linear Models (GLM's)

Transformations may normalize distributions, but they force the data into awkward scales for interpretation and do not always deal appropriately with assumption violations. Alternatively, separate functions to allow for non-linearity and heterogeneous variances can be used (Hastie and Tibshirani, 1990). Generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) deal with these issues in a natural way by using re-parameterization to induce linearity and by allowing a non-constant variance to be directly incorporated into the analysis. Hastie and Tibshirani (1990) suggest that this is closer to a re-parameterization of the model than a re-expression of the response. The link function describes the relationship between the mean and the linear predictor. A variance function relates the variance to the mean. The concept of generalized linear models (McCullagh and Nelder, 1989) provides an integrated approach that treats linear regression, anova, ancova, categorical and nominal models (logistic, probit, logit, log-linear) as special cases of a common family (O'Brien, 1992). The methods:

- share a common mathematical notation;
- link the response to predictors using a linear, additive structure; and
- draw the error component from a common family of probability distributions known as the exponential family.

A GLM is equivalent to the classical linear model when the link function is identity (1) and the error distribution normal. For classical regression models, residuals are used to assess the importance and relationship of a term in the model as well as to search for anomalous observations. For generalized models, residuals are additionally used to assess and verify the form of the variance as a function of the mean response (McCullagh and Nelder, 1989). By fitting data with error distributions that may be normal, binomial, Poisson, gamma, inverse gamma and others, GLM's dramatically extend the kind of data that may be modelled using interpretable regression methods. An important advantage of the GLM methods for soil-landscape modelling

is the ability to use various types of attribute data (e.g. nominal, ordinal, interval, ratio-scale) as predictors or response. McCullagh and Nelder (1989) indicate that an important characteristic of generalized linear models is that they assume independent (or at least uncorrelated) residuals.

2.5.4 Generalized Additive Models (GAM's)

The primary restriction of a generalized linear model is that it must be a linear function of the parameters of the model. Generalized additive models (Hastie and Tibshirani, 1986; 1990) subsume and extend GLM's by including fits of non-parametric functions to estimate the relationship between the response and explanatory variables. Non-parametric functions are estimated using data smoothing techniques and are attractive because they rely on the data to specify the form of the model (Cleveland, 1993). Hastie and Tibshirani (1990) state: "let the data show us the appropriate functional form. The idea behind a scatterplot smoother is to expose the functional dependence without imposing a rigid parametric assumption about that dependence." The term additive means that the model is a sum of terms. Some terms may be non-parametric, others linear and others a function of more than one predictor (multivariate).

A smoother summarizes the trend of a response variable as a function of one or more predictor measurements (Hastie and Tibshirani, 1990). The estimate of the trend is less variable than the reponse, hence the name smoother. No assumptions are made about the form of the dependence and so most smoothers are considered non-parametric. Hastie and Tibshirani (1990) indicate that smoothers have two main uses. The first, is description via methods discussed previously for EDA (Cleveland, 1993). The second, is to estimate the dependence of the mean of a response on the predictors, and thus serve as a building block for the estimation of additive models. A large variety of methods exist for data smoothing and include parametric techniques such as: natural cubic splines, B-splines, polynomials. Non-parametric techniques include: local regression (loess) and smoothing splines. The degrees of freedom consumed by

the non-parametric methods may be fractional depending on the functional form of the fit to the data. Many of these techniques are compared and contrasted by Hastie and Tibshirani (1990) and the literature suggests that differences are often small when appropriate smoothing parameters are chosen (Silverman, 1984; 1985; Muller, 1987). The two main questions in selection of a smoother are how to average the response values in each neighborhood and how big to take the neighborhoods. The various methods do this in different ways based on adjustable smoothing parameters which govern a fundamental tradeoff between bias and variance.

The deviance, or likelihood ratio plays the role of the residual sum of squares for generalized models, and can be used for assessing goodness-of-fit and for comparing models (Hastie and Tibshirani, 1990). Analysis of deviance may be performed in a similar manner to traditional analysis of variance. A dispersion parameter is used to calculate the deviance. For a model with a continuous response and normal errors, the dispersion parameter is equivalent to the variance and the deviance becomes equivalent to the residual sum of squares (McCullagn and Nelder, 1989). As with the residual sum of squares, the deviance can be made arbitrarily small by choosing an exact interpolating solution. But as stated above, the aim is to strike a balance between the number of parameters requiring estimation and the simplicity and applicability of the model. This is indicated by the degrees of freedom required or consumed for any particular model fit. An objective method of evaluating and comparing predictor variables is the Akaike Information Criterion (Akaike, 1974). This enables a stepwise selection of variables by comparing a statistic computed using residual deviance penalized by the number of parameters requiring estimation in a model fit.

Statistical Sciences (1993) state that additive models stumble when there are interactions among the various predictor terms and that local regression models provide much greater flexibility in that the model is fitted as a single smooth function of all the predictors. Hence, local regression smoothers impose no restrictions on the

relationships among the predictors. Other limitations include difficulties for casual users in understanding the complex number of smoothing parameters that may be tweaked for scatterplot smoothers. These may at times produce attractive curves for noisy data or data sparse areas in attribute space. Because smoothing methods rely on sample data to indicate the form of the function to use, a representative sample covering the overall variance of a soil variable (**s**) within a defined physiographic domain ($\boldsymbol{i}_{\mathbf{n}}$) is essential.

An advantage of the family of generalized linear and generalized additive modelling techniques is that an identical sequence of analysis methods may be used to select terms and change residual error models to improve the fit without transforming the response.

2.5.5 Geostatistical Models

Geostatistics or spatial statistics has developed as a specialized field because of the existence of spatial dependence in regionalized variables that are commonly of interest in the earth sciences (Webster and Oliver, 1990). The initial theory is attributed to Matheron (1965; 1971) and further empirical development by Krige (1966). The term kriging is used to refer to the spatial interpolation technique developed by Krige. In its simplest form, kriging is a method of weighted averaging within a neighborhood around observed values (Webster and Oliver 1990). A kriged value is a local estimate, and its goodness of fit depends on there being a number of measured values of the variable of interest close to the place for which the estimate is required. However, many variants of kriging (e.g. universal, block, punctual, co-) have developed since the work of Matheron and Krige (Cressie, 1991). All the methods depend on estimation of a variogram which requires a large number of sampling points at varied spacings (e.g. > 100) for robust estimation (Webster and Oliver, 1990).

When samples from two or more interdependent regionalized variables are available, cross semi-variograms (McBratney and Webster, 1983) may be produced to express the spatial relationships among variables. Co-kriging (Journal and Huijbregts, 1978; McBratney and Webster, 1983) may then be used to interpolate the values of a variable from measurements of it plus data on one or more other properties that have been more intensively sampled. Approaches that integrate regression and kriging have also been reported (Delhomme, 1979; Ahmed and DeMarsily, 1987). Odeh *et al.* (1994a; 1994b) compared prediction performance using regression, kriging, co-kriging and integrated regression-kriging techniques. They concluded that the precision and bias of prediction are dependent on the soil variable being predicted. In general, the regression-kriging methods performed better than the other techniques, but required the estimation of more parameters. They state: "In applying the prediction methods in the wider sense, the cost-benefit performance may determine the best method, i.e., whether an increase in precision (of prediction) is more than compensated by the cost of analysis (computing charge and time) and additional sampling." In the case of many general purpose soil survey programs, the intensity of spatial sampling is prohibitive.

Hastie and Tibshirani (1990) and Hutchinson and Gessler (1994) indicate that the formal equivalence between smoothing splines and kriging is well known (Matheron, 1981; Dubrule, 1983; 1984; Watson, 1984; Wahba, 1990). Hutchinson and Gessler (1994) indicate that spline methods do not require estimation of the variogram, yet, may provide similar information about the structure of the spatial variation and overall prediction error in a more robust way using generalized cross validation (GCV).

2.5.6 Tree Based Models

Tree based models (Breiman *et al.* 1984) are an alternative, non-parametric technique that uses recursive partitioning of a learning sample into increasingly homogeneous subsets. They may be used for both classification (e.g. prediction of

nominal and ordinal attributes) and regression (prediction of interval and ratio-scale attributes) and, as mentioned above, as an exploratory technique for uncovering structure and complex relationships in data (Breiman *et al.* 1984; Clark and Pregibon, 1992).

Clark and Pregibon (1992) indicate that tree based models are increasingly used for:

- devising prediction rules that can be rapidly and repeatedly evaluated;
- screening variables;
- assessing the adequacy of linear models; and
- summarizing large multivariate datasets.

They further state that in comparison to linear and additive models, tree-based models have the following advantages:

- easier to interpret when the predictors are a mix of nominal and continuous variables;
- invariant to transformations of predictor variables;
- more satisfactorily treat missing values;
- more adept at capturing non-additive behavior;
- detect more general interactions between predictor variables; and
- can model nominal or ordinal response variables with more than two levels.

Tree based modelling is effective when there is a significant interaction structure in the predictors because some attributes are conditionally important to only a subset of the population (Hastie and Tibshirani,1990). Likewise, these methods are particularly useful for geographic data because branching relationships are context sensitive and create sub-models which may relate to different parts of geographic space. For example, if soil variables on the upper parts of hillslopes in a catchment system exhibit strong relationships with predictor digital terrain attributes but

relationships change on descendence of the hillslope according to changes in landscape processes, Tree methods may capture this variation and partition the response learning set accordingly and develop different rules for each part of the landscape. Tree-based methods are also robust with respect to outliers and will by definition place them at an edge or terminal node of the tree (Walker and Coops, 1994).

Hastie and Tibshirani (1990) indicate that the piecewise-stepped nature of a regression tree surface is unattractive, and can be extremely inefficient for predicting a continuous response variable if the underlying surface is smooth. Tree-based prediction is probably most appropriate for classification of ordinal and nominal response data (Hastie and Tibshirani, 1990; Walker and Coops, 1994). The principal advantage of binary tree representations are the ease of interpretation for non-statisticians. The methods have many parrallels with the techniques used by soil mappers in integrating information when delineating map units. Pedologists can easily understand the visual representation of branching decisions and interpret process relationships.

2.5.7 Bayesian and Neural Network Modelling Techniques

Two approaches useful in situations similar to those where decision trees are appropriate are Bayesian analysis and neural networks. Bayesian statistical inference is a mathematical method used for decision making under conditions of uncertainty (Aspinall, 1992). The Bayesian decision model (Bayes' theorem) forms a framework for combining relative values of being right or wrong (subjective probabilities) with the probabilities of being right or wrong (conditional probabilities) (Aspinall and Hill, 1983). The appeal of this method is that it emulates the decision making process of experts in the field (Aspinall, 1992) and attempts to capitalize on accumulated knowledge for making predictions (O'Brien, 1992). Prior information or existing knowledge is used to estimate prior probabilities. These priors are then modified by experimentation or data into posterior probabilities or probabilities based on validation. These methods assume that subjective prior probabilities adequately represent the

uncertainty and that the predictor data sets are conditionally independent. O'Brien (1992) states that their is considerable argument about how to calibrate the prior terms (Bishop *et al.* 1975; Wrigley, 1985).

Neural networks are termed a "model free" approach to modelling that use supervised learning (Allman, 1989; Kosko, 1992). These methods do not consider explicitly the internal organization and spatial variability of data (Kacewicz, 1994). The feedforward backpropagation technique (Rumelhart *et al.* 1986) attempts to estimate an unknown function from observed random variables by minimizing an unknown expected error function. Input data (input nodes) or potential predictor variables are fed into the algorithm along with desired output (output nodes) or response variables (Skidmore and Turner, 1995). The algorithm then computes an internal pattern of weights for the middle layer(s) nodes of the network so that the input pattern produces the desired outputs at the output nodes. Most algorithms use an iterative process of modifying the weights to find the single set of weights that will solve all the input-output pairs used to train the network. Once the network is trained, it is possible to develop or solve outputs for unknown inputs.

Kacewicz (1994) indicated that it is very easy to teach the network with wrong, noisy or non-local information. Skidmore and Turner (1995) report disappointing results and conclude that the neural network backpropagation technique will not become a significant classification and analysis tool for GIS and remotely sensed data, except where relationships are obvious in the data set. Kacewicz (1994) cautions that the methods should not be used blindly and that users with a good understanding of the spatial distribution and characteristics of the data will do much better than uneducated users. He states that this is in contradiction to the concept of the methods being a "model free" approach. Sarle (1994) submits that neural networks are nothing more than nonlinear regression and discriminant analysis using new jargon.

2.6 SPATIAL PREDICTION

A principal application of a soil-landscape model is to predict, with a stated accuracy and precision, the value of the response attribute at un-sampled locations. Austin and McKenzie (1988) divide spatial extension or prediction techniques into two broad groups of (i) interpolative and surface fitting methods, and (ii) environmental correlation methods. Discussion of both groups are encompassed in previous sections and, as indicated, many techniques that bridge the gap between these two categories have been reported in the literature (Stein and Corsten, 1991; Odeh *et al.* 1994). Interpolative and surface fitting methods involve the weighting of neighboring observations to estimate the value of a variable at an intervening point. These techniques require a large number of sample observations distributed throughout the geographic space in the vicinity of the required predictions. Data smoothers and kriging techniques that use only the sample observations fall into this general category.

Austin and McKenzie (1988) state that environmental correlation methods are a closer quantitative analogue of traditional survey methods. This approach will be of most value when useful predictor environmental variables can be more easily measured and are available in a spatially continous way (Austin and McKenzie, 1988; Gessler *et al.* 1995). Spatial prediction using GLM, GAM, tree based and other modelling techniques with continuous predictor environmental variables is achieved with a GIS using map algebra to compute predictive equations on digital overlays to produce a predictive end product. These techniques enable digital spatial analysis to become an integral tool used throughout the process outlined in this review (Tomlin, 1983; 1990; Burrough, 1986; Berry, 1993).

2.7 SUMMARY

The previous sections have provided a review of developing concepts and techniques that may improve soil resource inventory. From this, several advances are possible:

- explicit modelling process - from setting the overall model context to spatial prediction of individual response soil attributes;
- use of quantitative and digital data, where possible, with defined measurement scales and levels of support;
- integration of landscape process understanding via environmental gradients with statistical and spatial theory to guide gathering of sample evidence;
- dynamic data exploration to analyze sample evidence and environmental correlations;
- opportunity to select appropriate models for the data and purpose;
- quantitative spatial predictions with confidence estimates; and
- the capacity for model testing and improvement.

Traditional and contemporary methods for developing general purpose soil resource inventories do not have these features. The approach discussed here places general purpose soil resource inventory on the scientific base advocated by (Hewitt, 1993). The concepts developed and literature reviewed provide the base for implementation of the approach suggested by McSweeney *et al.* (1994). The following chapters will demonstrate an implementation with data from three study areas in southeastern Australia.

2.8 REFERENCES CITED

Allen, T.F.H., and T.W. Hoekstra. 1992. Toward a unified ecology. Columbia Univ. Press, NewYork.

Allman, W.F. 1989. Apprentices of wonder: inside the neural network revolution. Bantam, New York.

Ahmed, S., and G. DeMarsily. 1987. Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. Water Resources Res. 23:1717-1737.

Akaike, H. 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19. 716-722.

Aspinall, R. 1992. An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. Int. J. Geog. Inf. Sys. 2:105-121.

Aspinall, R., and A.R. Hill. 1983. Clinical inferences and decisions: I. Diagnosis and Bayes' theorem. Opthalmic Physiological Optics. 3:295-304.

Austin, M.P. and N.J. McKenzie. 1988. Data analysis. pp.210-231. *In* R.H. Gunn, J.A. Beattie, R.E. Reid and R.H.M. van de Graaff (eds.) Australian Soil and Land Survey Handbook. Inkata, Melbourne.

Austin, M.P., and P.C. Heyligers. 1989. Vegetation survey design for conservation: gradsect sampling of forests in northeastern New South Wales. Biol. Conserv. 50:13-32.

Becker, R.A. and W.S. Cleveland. 1987. Brushing scatterplots. Technometrics 29:127-142.

Beckett, P.H.T. 1968. Method and scale of land resource surveys, in relation to precision and cost. p.51-63. *In* G.A. Stewart (ed.) Land Evaluation. Macmillan, Australia.

Beckett, P.H.T., and R. Webster. 1971. Soil variability: a review. Soils Fert. 34:1-15.

Beckett, P.H.T., and S.W. Bie. 1978. Use of soil and land system maps to provide soil information in Australia. CSIRO Aust. Div. Soils Tech. Paper No. 33. Adelaide, Australia.

Bell, J.C. 1990. A GIS-based soil-landscape modelling approach to predict soil drainage class. Ph.D. Dissertation. The Pennsylvania State University. University Park, PA.

Bell, J.C., R.L. Cunningham, and M.W. Havens. 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. Soil Sci. Soc. Am. J. 56:1860-1866.

Bell, J.C., R.L. Cunningham, and M.W. Havens. 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Sci. Soc. Am. J. 58:464-470.

Berry, J.K. 1993. Cartographic modelling: the analytical capabilities of GIS. p.58-74. *In* M.F. Goodchild, B.O. Parks and L.T. Steyaert (eds.) Environmental Modelling with GIS. Oxford University Press, New York.

Bevan, K.J., and M.J. Kirkby. 1979. A physically-based variable contributing area model of basin hydrology. Hydro. Sci. Bull. 24:43-69.

Bie, S.W., and Beckett, P.H.T. 1970. The costs of soil survey. Soils Fert. 33:203-217.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, Mass.

Bouma, J. 1983. Use of soil survey data to select measurement techniques for hydraulic conductivity. Agric. Water Manag. 6(2/3):177-190.

Bouma, J. 1989. Land qualities in space and time. p.3-13 In J. Bouma and A.K. Bregt (eds.) Land Qualities in Space and Time. Pudoc, Wageningen, The Netherlands.

Box, G.E.P., W.G. Hunter, and J.S. Hunter. 1978. Statistics for experimenters. Wiley, New York.

Breiman, L., J.H. Friedman, R. Olshen, and C.J. Stone. 1984. Classification and regression trees. Wadsworth International Group, Belmont, California.

Burrough, P.A. 1986. Principles of geographical information systems. Clarendon Press, Oxford.

Butler, B.E. 1956. Parna - an aeolian clay. Australian J. of Sci. 18:145-151.

Butler, B.E. 1964. Can pedology be rationalized? A review of the general study of soils. Australian Soc. Soil Sci. Pub. No. 3. Adelaide, Australia.

Butler, B.E. 1980. Soil classification for soil survey. Clarendon Press, Oxford.

Butler, B.E. 1982. A new system for soil studies. J. Soil Sci. 33:581-595.

Chambers, J.M. and Hastie, T.J. 1992. *Statistical models in S.* Wadsworth & Brooks, Los Angeles.

Clark, L.A. and D. Pregibon. 1992. Tree based models. p. 377-419. *In* Statistical Models in S. Wadsworth & Brooks, Pacific Grove, CA.

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, New Jersey.

Cleveland, W.S. and M.E. McGill. 1988. Dynamic graphics for statistics. Chapman and Hall, New York.

Cook, R.D., and S. Weisberg. 1982. Residuals and influence in regression. Chapman and Hall, New York.

Cressie, N.A.C. 1991. Statistics for spatial data. Wiley and Sons, New York.

Curtis, J.T. 1959. The vegetation of Wiscsonsin: an ordination of plant communities. University of Wisconsin Press, Madison.

Daniels, R.B., and R.D. Hammer. 1992. Soil geomorphology. Wiley and Sons, New York.

Delhomme, J.P. 1979. Spatial variability and uncertainty in groundwater flow parameters: a geostatistical approach. Water Resourc. Res. 15:269-280.

Dijkerman, J.C. 1974. Pedology as a science: the role of data, models and theories in the study of natural systems. Geoderma. 11:73-93.

Draper, N.D., and H. Smith. 1981. Applied regression analysis. 2nd Ed. Wiley and

Sons, New York.

Dubrule, O. 1983. Two methods with different objectives: splines and kriging. Math. Geol. 15:687-699.

Dubrule, O. 1984. Comparing splines with kriging. Comput. Geosci. 10:327-338.

Evans, I.S. 1972. General geomorphometry, derivatives of altitude, and descriptive statistics. *In* R.J. Chorley (ed.) Spatial analysis in geomorphology. Methuen, London.

Fisher, R.A. 1958. Statistical methods for research workers. Hafner, New York.

Fisherkeller, M.A., J.H. Friedman, and J.H. Tukey. 1988. Prim-9: An interactive multidimensional data display and analysis system. p. 91-109. *In* W.S. Cleveland and M.E. McGill (eds.) Dynamic Graphics for Statistics. Chapman and Hall, New York.

Fitzpatrick, E.A. 1967. Soil nomenclature and classification. Geoderma. 1:91-105.

Fitzpatrick, E.A. 1993. Soil horizons. Catena 20:1-5.

Gerrard, A.J. 1990. Soil variations on hillslopes in humid temperate climates. Geomorphology 3:225-244.

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modelling and the spatial prediction of soil attributes. Int. J. of Geog. Inf. Sys. 9(4):421-432.

Gibbons, F.R. 1961. Some misconceptions about what soil surveys can do. J. Soil Sci. 12:96-100.

Gillison, A.N., and K.R.W. Brewer. 1985. The use of gradient directed transects or gradsects in natural resource surveys. J. Environ. Management 20:103-127.

Hastie, T., and R. Tibshirani. 1986. Generalized additive models. Statist. Sci. 1:297-318.

Hastie, T., and R. Tibshirani. 1990. Generalized additive models. Chapman and Hall, New York.

Hewitt, A.E., 1993. Predictive modelling in soil survey. Soils Fert. 3:305-314.

Hole, F.D. and J.B. Campbell. 1985. Soil landscape analysis. Rowman and Allenheld, Totowa.

Hoosbeek, M.R., and R.B. Bryant. 1992. Towards the quantitative modeling of pedogenesis - a review. Geoderma. 55:183-210.

Hutchinson, M.F., and P.E. Gessler. 1994. Splines - more than just a smooth interpolator. Geoderma 62:45-67.

Jenny, H. 1941. Factors of soil formation: a system of quantitative pedology. McGraw-Hill, New York.

Jenny, H. 1980. The soil resource: origin and behavior. Springer-Verlag, New York.

Journal, A.G., and C.J. Huijbregts. 1978. Mining geostatistics. Academic Press, London.

Kacewicz, M. 1994. Model-free estimation of fracture aperture with neural networks. Math. Geol. 26(8):985-994.

Kosko, B. 1992. Neural networks and fuzzy systems - a dynamical systems approach to machine intelligence. Prentice Hall, New York.

Krige, D.G. 1966. Two-dimensional weighted moving average trend surfaces for ore evaluation. Journal South African Institute Mining Metallurgy. 66:13-38.

Lauren, J.G., R.J. Wagenet, J. Bouma, and J.H.M. Wosten. 1988. Variability of saturated hydraulic conductivity in a Glossaquic Hapludalf with macropores. Soil Sci. 145(1):20-28.

Matheron, G. 1965. Les variables regionalisees at leur estimation. Masson, Paris.

Matheron, G. 1971. The theory of regionalized variables and its applications. Cahiers du Centre de Morphologie Mathematique de Fontainebleau, no. 1.

Matheron, G. 1981. Splines and kriging: their formal equivalence. Syracuse Univ. Geol. Contrib. 8:77-95.

McBratney, A.B. 1993. Some remarks on soil horizon classes. Catena. 20:427-430.

McBratney, A.B., and R. Webster. 1983. Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. J. Soil Sci. 34:137-162.

McCullagh, P., and J.A. Nelder. 1989. Generalized linear models. Second Ed. Chapman and Hall, London.

McKenzie, N.J. 1991. A strategy for coordinating soil survey and land evaluation in Australia. CSIRO Div. of Soils. Divisional Report 114. Adelaide, Australia.

McKenzie, N.J., and M. Austin. 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329-355.

McSweeney, K., P.E. Gessler, B. Slater, R.D. Hammer, J. Bell, and G.W. Petersen. 1994. Towards a new framework for modelling the soil-landscape continuum. p.127-145. *In* Factors of soil formation: a fiftieth anniversary retrospective. SSSA Special Pub. 33. Madison, WI.

Moore, I.D., A.R. Ladson, and R. Grayson 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydro. Proc. 5:3-30.

Moore, I.D., P.E. Gessler, G.A. Neilsen, and G.A. Peterson. 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57:443-452.

Milne, G. 1935. Some suggested units of classification and mapping particularly for East African soils. Soil Research 4:3.

Muller, H.G. 1987. Weighted local regression and kernel methods for nonparametric curve fitting. J. Am. Statist. Assoc. 82:231-238.

Nelder, J.A., and R.W.M. Wedderburn. 1972. Generalised linear models. J. R. Statist. Soc. A. 135:370-384.

Odeh, I.O.A. 1990. Soil pattern recognition in a South Australian subcatchment. Ph.D. thesis. University of Adelaide.

Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1994a. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63:197-214.

Odeh, I.O.A., P.E. Gessler, N.J. McKenzie, and A.B. McBratney. 1994b. Using attributes derived from digital elevation models for spatial prediction of soil properties. p. 451-463. *In* Proceedings of Resource Technology 94. 26-30 Sept. 1994, Melbourne, Australia. University of Melbourne, Melbourne.

O'Brien, L. 1992. Introducing quantitative geography. Routledge:London.

O'Loughlin, E.M. 1986. Prediction of surface saturation zones in natural catchments by topographic analysis. Wat. Res. Res. 22(5):794-804.

Price, A.G. 1994. Measurement and variability of physical properties and soil water distribution in a forest podzol. J. Hydrol. 161:347-364.

Reid, R.E. 1988. Soil survey specifications. p.60-72. *In* R.H. Gunn, J.A. Beattie, R.E. Reid and R.H.M. van de Graaff (eds.) Australian Soil and Land Survey Handbook. Inkata, Melbourne.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams. 1986. Learning internal representations by error propagation. p318. *In* J.L. McClelland and D.E. Rumelhart (eds.) Parallel distributed processing: explorations in the microstructure of cognition. MIT Press, Cambridge, MA.

Silverman, B.W. 1984. A fast and efficient cross-validation method for smoothing parameter choice in spline regression. J. Am. Statist. Assoc. 79:584-9.

Silverman, B.W. 1985. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. J.R. Statist. Soc. B. 47:1-52.

Simonson, R.W. 1959. Outline of a generalized theory of soil genesis. Soil Sci. Soc. Am. Proc. 23:152-156.

Skidmore, A.K., and B.J. Turner. 1995. Remote sensing and geographical information systems as forestry tools: a critique. p. 41-48. *In* Proc. Institute of Forestors of Australia 16th Biennial Conference. Canberra, Australia.

Slater, B.K. 1994. Continuous classification and visualization of soil layers: a soil-landscape model of Pleasant Valley Wisconsin. Ph.D. dissertation. University of Wisconsin, Madison, WI.

Slater, B.K., K. McSweeney, S.J. Ventura, B.J. Irvin, and A.B. McBratney. 1994. A spatial framework for integrating soil-landscape and pedogenic models. p. 169-185. *In* Quantitative Modeling of Soil Forming Processes. SSSA Special Publication 39. Madison, WI.

Snedecor, G.W., and W.G. Cochran. 1980. Statistical Methods. Iowa State Univ. Press, Ames, IA.

Snee, R.D. 1985. Experimenting with a large number of variables. p. 25-35. *In* R.D. Snee (ed.) Experiments in Industry. American Society of Quality Control. Milwaukee, WI.

Speight, J.G. 1968. Parametric description of landform. p239-250. *In* G.A. Stewart (ed.) Land Evaluation. Macmillan, Melbourne.

Speight, J.G. 1974. A parametric approach to landform regions. Special Publ. no. 7. Institue of British Geographers.

Statistical Sciences. 1993. S-PLUS Guide to statistical and mathematical analysis. Version 3.2. StatSci, a Division of MathSoft, Inc. Seattle, WA.

Stein, A. and L.C.A. Corsten. 1991. Universal kriging and cokriging as a regression procedure. Biometrics 47:575-587.

Steur, G.G.L. 1961. Methods of soil surveying in use at the Netherlands Soil Survey Institute. Boor en Spade 11:59-77.

Stigler, S.M. 1981. Gauss and the invention of least squares. Annals Statistics. 9:465-474.

Stigler, S.M. 1986. The history of statistics. Belknap Press, Cambridge, U.K.

Tomlin, C.D. 1983. A map algebra. *In* Proc. Harvard Computer Conf. 31 July-4 Aug. 1983. Cambridge, Mass.

Tomlin, C.D. 1990. Geographic information systems and cartographic modelling. Prentice Hall, Englewood Cliffs, NJ.

Troeh, F.R. 1964. Landform parameters correlated to soil drainage. Soil Sci. Soc. Am. J. 32:102-104.

Tukey, P.A. and J.W. Tukey. 1981. Graphical display of data sets in 3 or more dimensions. p.189-275. *In* V. Barnett (ed.) Interpreting Multivariate Data. Wiley, Chichester, U.K.

Tukey, J.W. 1977. Exploratory data analysis. Addison-Wesley: Reading, Mass.

van Wesenbeeck, I.J., and R.G. Kachanoski. 1994. Effect of variable horizon thickness on solute transport. Soil Sci. Soc. Am. J. 58:1307-1316.

Vitousek, P. 1994. Factors controlling ecosystem structure and function. p.84-101. *In* Factors of soil formation: a fiftieth anniversary retrospective. SSSA Special Pub. 33. Madison, WI.

Wahba, G. 1990. Spline models for observational data. CBMS-NSF Regional Conf. Series in Applied Mathematics. SIAM, Philadelphia.

Walker, P.A., and N.C. Coops. 1994. Classification trees. *In* R. Aspinall and P.A. Walker (eds.) Workshop evaluating spatial modelling techniques. Monograph. Macaulay Land Use Research Institute, Aberdeen.

Walker, P.H., G.F. Hall, and R. Protz. 1968. Relation between landform parameters and soil properties. Soil Sci. Soc. Am. J. 32:102-104.

Watson, G.S. 1984. Smoothing and interpolation by kriging and with splines. Math. Geol. 16:601-615.

Watson, J.P. Soil catenas. 1965. Soils Fert. 28:307-310.

Webster, R., and M.A. Oliver. 1990. Statistical methods in soil and land resource survey. Oxford University Press, Oxford.

Weisberg, S. 1980. Applied linear regression. Wiley and Sons, New York.

Wilk, M.B. and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. Biometrika 55:1-17.

Wrigley, N. 1985. Categorical data analysis for geographers and environmental scientists. Longman, London.

Yates, F. 1937. The design and analysis of factorial experiments. Imperial Bureau Soil Science Technical Committee. 35:4-95.

Chapter Three: Sampling and Model Development

3.1 INTRODUCTION

3.1.1 Hypothesis and Concepts

The aim of this chapter is to test the mechanics of a sampling and model development approach based on the ideas developed in Chapter Two. The hypothesis is that:

- explicit and quantitative environmental correlations can be derived to spatially predict individual soil attributes using statistical models with stated levels of uncertainty and model complexity.

The approach uses an upper meso-scale environmental stratification to delineate three study areas and a provisional model based on a lower meso-scale digital terrain attribute environmental gradient to allocate field sample locations for measurement of a range of soil attributes over each study area. A subset of the data collected from the Ladysmith study area is used in this chapter to test the hypothesis.

Subsequent chapters will use the models and techniques demonstrated here to evaluate environmental correlations in more detail and develop preliminary soil-landscape process interpretations. The focus here is on the mechanics of a new approach that may be applied for general purpose natural resource inventory.

3.2 MATERIAL AND METHODS

3.2.1 Study Region

The study region is bounded by the Wagga Wagga and Tarcutta 1:100 000 topographic map sheets (147° 00', 35°00'; 148°00', 35°00'; 148°00', 35°30'; 147°00', 35°30') on the southwest slopes of the Great Dividing Range in southeastern New South Wales (Figure 3.1). This region was chosen because of its diverse range of

geology, climate, landforms and land uses. It is typical of large parts of the arable portion of the Murray-Darling Basin.

GIS Development

A regional GIS was developed to hold digital datasets on geology, climate, gamma radiometrics, soils, terrain, hydrography, land ownership, geodetic control, transportation (e.g. roads, railroads), satellite imagery, orthophotography and collected field sample data (Gessler and Ashton, in prep.). At the upper meso-scales (1:100 000 cartographic scale), nominal attribute maps of bedrock geology (Raymond, 1992) and soils (Chen and McKane, in press) were digitized. A regional digital elevation model provided by SPOT was used with the ANUCLIM software (McMahon *et al.* 1995) to produce several climatic surfaces (e.g. total annual precipitation, precipitation of the dryest quarter, mean annual temperature, annual mean radiation) on a grid node spacing of approximately 245m.

At the mid meso-scales, the Australian Geological Survey Organization provided gamma radiometric data at a grid node spacing of 50m. Airborne gamma-spectrometry provides spatial images of the geochemistry of the top 0.30-0.45m of the rock or soil layer by measuring the abundance of gamma-rays produced by the radioactive decay of potassium, thorium and uranium along with a total count image of all gamma-rays sensed (Bierwirth *et al.* 1996).

At the lower meso-scales, digital data from twelve 1:25 000 topographic map sheets, covering the entire Wagga Wagga 1:100 000 sheet and the western half of the Tarcutta 1:100 000 sheet, were supplied by the New South Wales, Land Information Centre. These included 10m elevation contours, streamlines and spot heights. Additional linework for land ownership boundaries and roads was also digitized from the 1:25 000 sheets. The contours, streamlines and spot heights were used as input to the ANUDEM software (Hutchinson, 1989) for generation of 20m grid node spacing digital elevation models. The TAPESG software (Moore, 1992; Gallant, 1996) was used to generate primary terrain attributes. Flow accumulation was modelled using

the TAPESG flow dispersion algorithm with a cross-over threshold of 100 grid nodes. This causes a change from multiple drainage direction dispersive flow to the deterministic eight direction (d8) flow technique that channels all the flow from one grid node to the next in the direction of steepest descent. Moore *et al.* (1993b) report that this approach is more physically realistic than methods that use only the d8 method because it allows dispersive flow in the uplands and channelized flow along stream channels.

3.2.2 Environmental Stratification and Study Area Selection

Environmental stratification (i_n) at the upper meso-scales for selection of study areas was based on two bedrock map units from the digital bedrock geology map (Raymond, 1992). This stratification assumes the geology map is accurate. Three rectangular study areas were chosen to coincide with extensive, gently rolling, upland areas of Ordovician metasediment and Silurian granite bedrock types common to the eastern slopes of the Murray-Darling Basin (Figure 3.1). Three areas were chosen to facilitate comparison of models and environmental relationships in the region. "Upland" areas or areas dominated by hillslope processes were chosen, more specifically, to test the potential utility of lower meso-scale (20m) digital terrain attributes as predictors for exploration and modelling of soil-landscape patterns. Areas on other parent materials as defined by the geology map were excluded because of the assumption that different soil-landscape models are required in these domains. The study area boundaries were further refined to encompass entire sub-catchments on each bedrock type to ensure proper computation of flow accumulation critical for geomorphometric and hydrological characterization.

The Brucedale study area is on Silurian granite with an aeolian clay or parna cover (Butler, 1956), is 8,526 hectares in size and has mixed agricultural land uses of cereal cropping and pastoral grazing. The Ladysmith and Griggward study areas are on Ordovician metasediments and cover 5,664 and 5,304 hectares, respectively, with pastoral grazing the dominant land use. Figure 3.1 shows the location, size and

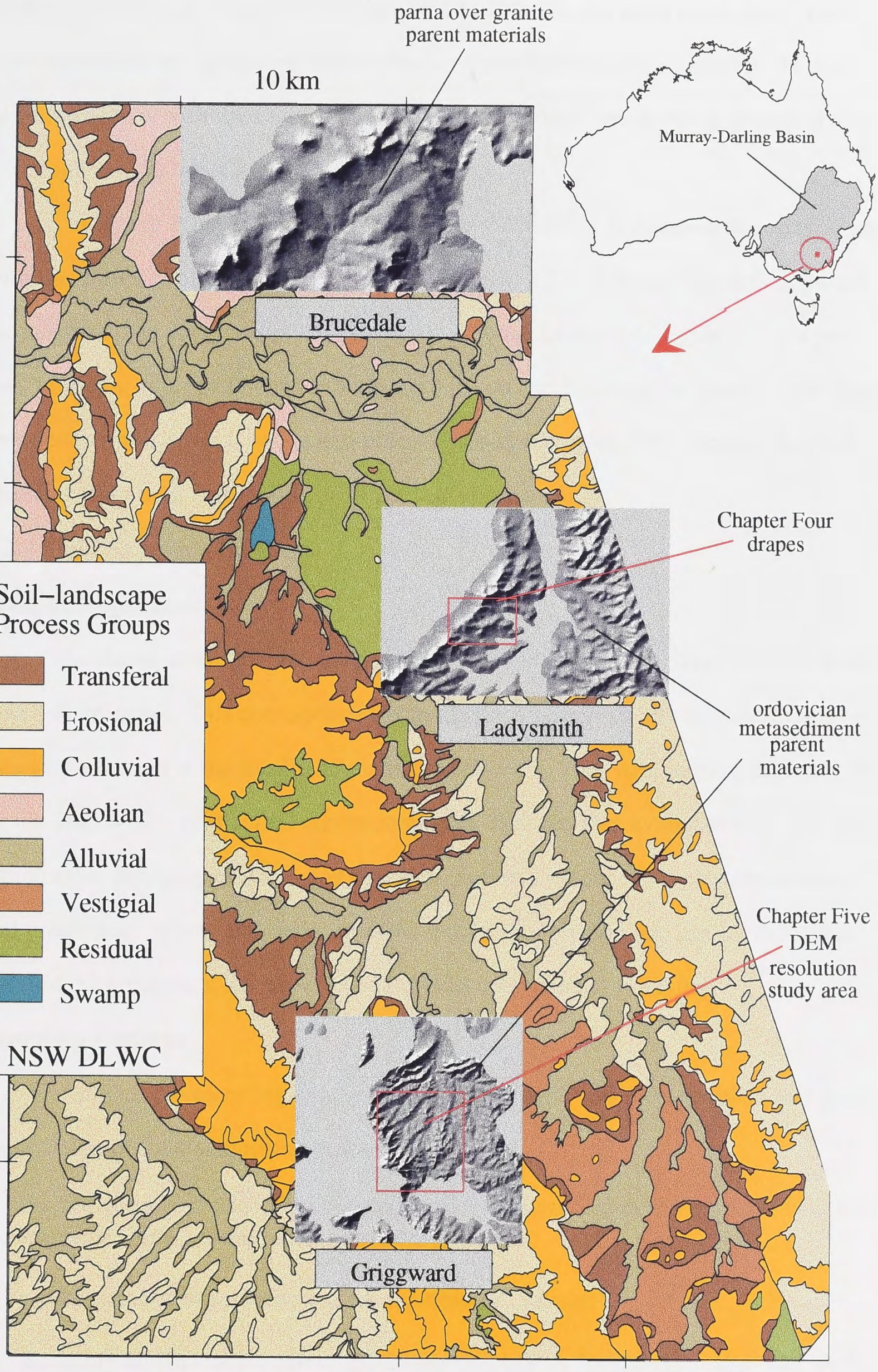


Figure 3.1 Study Area DEM Hillshades

landform patterns as visualized through DEM hillshades for each study area. The hillshades indicate significant differences between the hillslope patterns of the two parent materials with the Ordovician metasediment hillslopes showing greater dissection, higher local relief and shorter slope lengths.

In Figure 3.1, the area encompassed by the red box in Ladysmith corresponds with the drapes shown in Figures 3.7, 3.8, 3.14 and 3.21. The red box at Griggward locates the study area used in Chapter Four. The background of figure 3.1 is a process grouping aggregation of soil map units (Chen and McKane, in press.). The study areas encompass several soil map units and subsequent work will compare derived products from both soil-landscape modelling approaches.

3.2.3 Field Sampling Strategy

Scale of Application

As stated above, the intended scale of application for developed models is the lower meso-scale. This corresponds to the hillslope within small catchments and provides information at the local land management level. This was, in part, determined by the 20m x 20m grid digital elevation models derived from available 1:25 000 scale topographic mapping in the region that form the base product for spatial extension and prediction. It was also determined by a desire to develop and test methods that provide information at a scale useable for land management planning.

Sampling Criteria

An iterative sampling strategy using four criteria was used to select and guide the location of field samples as follows:

- the sampling plan used a provisional model for stratified random sampling at even intervals along a terrain environmental gradient;
- randomization was used to ensure an unbiased sample;
- sampling inefficiencies due to spatial dependence in the provisional model were minimized and hence, optimized in geographic space; and
- locational error between the provisional (digital) model and the real world sample locations was minimized.

Details of each aspect are discussed below.

Provisional Model and Attribute Space Stratification

A common provisional model is the catena (Latin = a chain) soil-landscape model (Milne 1935) that implies a concordance of soil pattern with landform as one traverses from hilltop to valley bottom down hillslopes. The compound topographic index (CTI), often referred to as the steady-state wetness index (Bevan and Kirkby, 1979; Moore *et al.* 1991), may be considered a quantification of the catenary landscape continuum. It is defined as:

$$CTI = \ln(A_s / \tan \beta) \quad (1)$$

where A_s is specific catchment area (area (m^2) per unit width orthogonal to the flow direction) and β is slope angle in degrees.

The CTI was used in this work as an explicit and quantitative provisional model to stratify and randomly sample each study area. The CTI probability density function for the geographic space of each study area quantifies the CTI attribute space and provides a hillslope environmental gradient for sampling. Figure 3.2 displays the CTI density functions for each study area. To spread samples evenly along this gradient, CTI attribute space was divided into five equal quantile classes (20th percentiles) for each study area. The selection of five classes was based on visual analysis of the patchiness of classes along various quantile segmentations. Figure 1 in Appendix Two shows a segmentation of CTI attribute space and visualization of spatial patch patterns using this approach for a sub-area of the Griggward study area. The spatial strata or patches defined by the quantile classes were used for randomization to meet the second sampling criterion.

Distribution in Geographic Space

No *a priori* information was available on the spatial dependence structure of any soil attributes in the study areas. Therefore, it was postulated that the spatial dependence structure of the CTI related in a general way to the spatial dependence

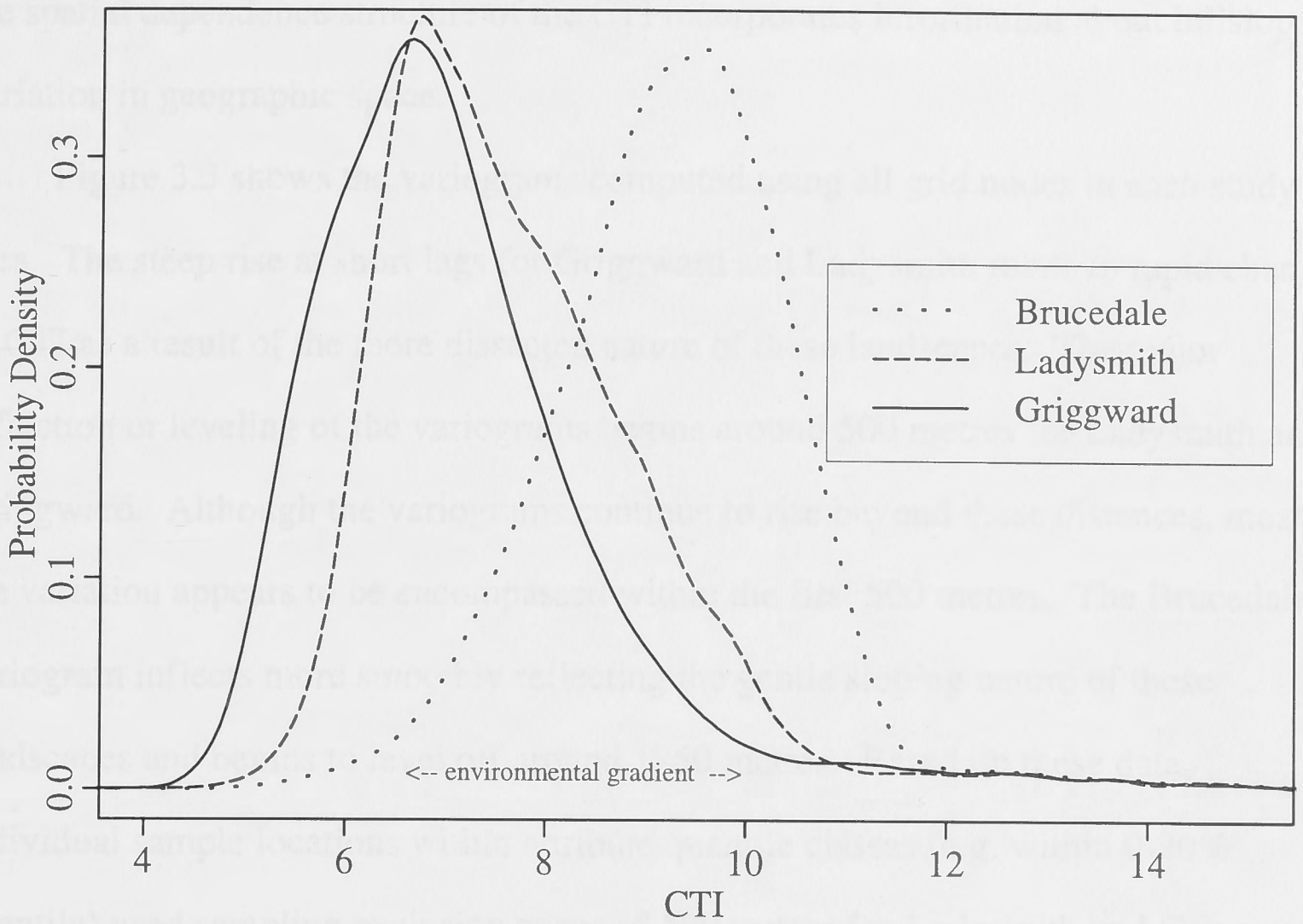


Figure 3.2 Study Area CTI Distributions (environmental attribute space)

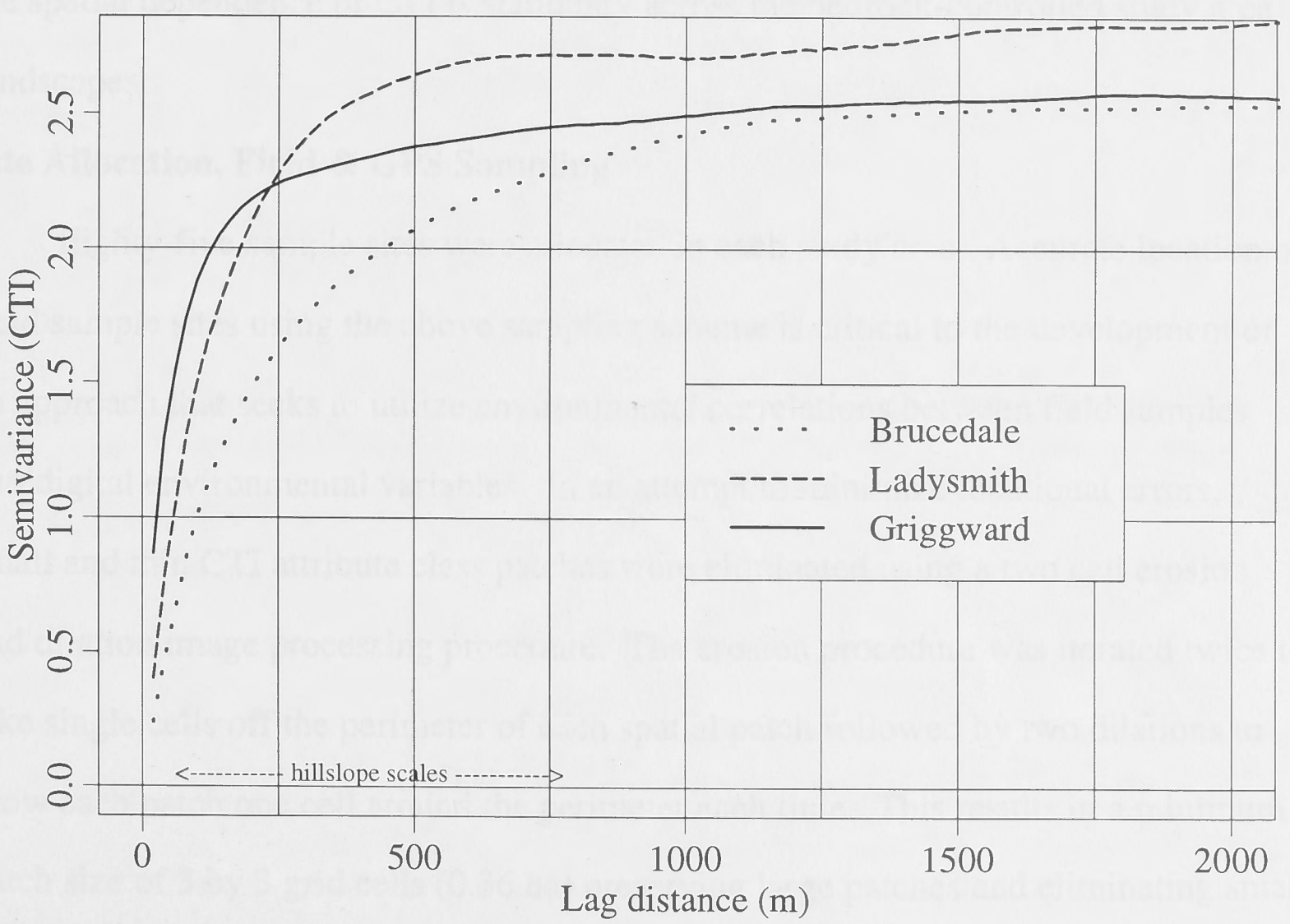


Figure 3.3 Study Area CTI Variograms (geographic space)

structure of a range of soil attributes. Spacing sample sites using a quantification of the spatial dependence structure of the CTI incorporates information about hillslope variation in geographic space.

Figure 3.3 shows the variograms computed using all grid nodes in each study area. The steep rise at short lags for Griggward and Ladysmith relate to rapid change in CTI as a result of the more dissected nature of these landscapes. The major inflection or leveling of the variograms begins around 500 metres for Ladysmith and Griggward. Although the variograms continue to rise beyond these distances, most of the variation appears to be encompassed within the first 500 metres. The Brucedale variogram inflects more smoothly reflecting the gentle sloping nature of these landscapes and begins to level off around 1250 metres. Based on these data, individual sample locations within attribute quantile classes (e.g. within 0-20% quantile) used sampling exclusion zones of 500 metres for Ladysmith and Griggward, and 1250 metres for Brucedale. This ensured that no two samples from the same quantile class were located in close proximity. An assumption of this approach is that the spatial dependence of CTI is stationary across the bedrock-controlled study area landscapes.

Site Allocation, Field & GPS Sampling

Eighty-five sample sites were allocated in each study area. Accurate location of field sample sites using the above sampling scheme is critical to the development of an approach that seeks to utilize environmental correlations between field samples and digital environmental variables. In an attempt to minimize locational errors, small and thin CTI attribute class patches were eliminated using a two cell erosion and dilation image processing procedure. The erosion procedure was iterated twice to take single cells off the perimeter of each spatial patch followed by two dilations to grow each patch one cell around the perimeter each time. This results in a minimum patch size of 3 by 3 grid cells (0.36 ha) preserving large patches and eliminating small

patches. This eliminates some areas from potential selection and adds some bias to the patch selection process.

Seventeen samples were distributed in each of the five CTI quantile classes according to the following iterative scheme. The patches for each class were numbered from 1 to *n* (total number of patches in the study area for the percentile class). A random number generator was used to produce a random number vector of length *n*. Patch numbers were selected sequentially from the vector and Australian Map Grid (AMG) coordinates produced for the approximate center of each patch. A visual display of sample sites with a range radius circle indicated if sites were within the exclusion zone (1250m - Brucedale, 500m - Ladysmith, 500m - Griggward) of previously selected sites for the same quantile. If so, they were discarded and the next random patch selected until seventeen sites were allocated for each of the five quantile classes. Slope direction or aspect was also displayed for each patch. If large patches were selected, aspects that had not previously been selected were searched for within the patch to also provide an approximate spreading of sample sites in aspect attribute space. This sample allocation process was done using the GIS.

Topographic maps displaying the sample sites and AMG coordinates were used with a hand-held Trimble Pathfinder Global Positioning Satellite (GPS) receiver to locale sample site locations in the field. The GPS manufacturer reports that the handheld receiver will provide locations accurate to within 10 metres (Trimble Navigation Ltd., 1992). Site values of the slope gradient, aspect, elevation and specific catchment area attributes computed from the DEM were also used to refine site placement and ensure a better match or consistency between the modelled CTI and real world geographic space (i.e. the actual field location). A wooden peg was placed in the ground at each site and revisited for sample collection.

A truck-mounted drill rig push tube was used to take a 71mm diameter soil core to a maximum depth of 2.3m or two drill tube lengths. The cores were packed in PVC sewer pipe for transport back to the laboratory for description and sampling.

Each sample site was located more precisely after sampling by occupying the site with a Trimble Pathfinder for five minutes coincident with data collection at a Trimble 4000 SSE Geodetic basestation. The data were differentially processed and corrected to provide a more accurate location. The GPS manufacturer reports that this process provides locations accurate to within one metre. The corrected coordinates were then used to re-sample all digital environmental variables in the GIS database for entry into a master database. This process ensured a better match between the field sample location and modelled environmental variables potentially useful as predictors. In the Griggward study area, two 4000 SSE basestations were used instead of a Pathfinder and 4000 SSE to provide better positioning as discussed below in Section 3.2.

Thirty-five percent of the sites (6 per class) for the Ladysmith and Brucedale areas were randomly selected for collection of duplicate cores taken 5m from the original sample location. These samples will be used in a subsequent study to evaluate short-range variation and the representativeness of soil core samples to the surrounding soil mantle. Handheld gamma ray spectrometry measurements (K, Th, U) for evaluation and correlation with airborne gamma radiometrics were also collected at the Ladysmith sites prior to soil core acquisition (Bierwirth *et al.* 1996).

3.2.4 Soil Core Description and Sampling for Lab Analyses

Basic soil profile morphology (e.g. texture, color, structure) for the soil cores was described according to McDonald *et al.* (1990) and used to derive horizon or soil layers for each core. The large number of soil cores (~ 500) made it impossible to sample even depth increments down the profile for lab analyses. Instead, samples were taken at the approximate depth centre of each horizon or layer. Soil material was taken from the volumetric centre (i.e. avoiding the outer core material) of each core, gently ground by mortar and pestle to break aggregates, air dryed, and passed through a 2mm sieve before placing in a sample vial. An approximate average volume and weight of each sample being 70ml and 80 grams after grinding. The

samples were used for chemical (Griggward, Ladysmith, Brucedale) and particle size analyses (Griggward, Brucedale).

3.2.5 Methods of Lab Analysis

Chemical Analyses

Laboratory measurements were made of: pH, total carbon, basic exchangeable cations (Ca, Mg, Na, K), cation exchange capacity and electrical conductivity. Soil pH was determined with a Radiometer pH meter using a combined glass/calomel electrode in a 1:5 soil water suspension after 1 hour of rotational shaking and 0.5 hour of settling. Total carbon (g/100g soil) was determined using a Leco CR-12 combustion furnace (Merry and Spouncer, 1988) with an infra-red CO_2 detector. Exchangeable sodium, potassium, calcium and magnesium (cmol/kg soil) were extracted with 0.01 M silver thiourea (Searle, 1986) and cation exchange capacity (CEC, cmol/kg soil) was determined by Ag remaining in silver thiourea extractant (Searle, 1986). Electrical conductivity (S m^{-1}) was measured using a Radiometer conductivity meter in a 1:5 soil water suspension.

Particle Size Analysis

Particle size analysis was performed for all samples from the Griggward and Brucedale study areas. Samples were prepared as described by Hutka and Ashton (1995). Particle size analyses were performed using a physical sieving process for particles greater than 53 μm and the less than 53 μm fraction determined by analysis with the Sedigraph 5100 Particle Size System (Hutka, 1994). This yielded quantitative data and a plot of the cumulative mass percent finer versus equivalent spherical diameter. Data were also provided on median and modal diameters on a mass distribution, number distribution and surface area distribution basis.

3.2.6 Exploratory Data Analysis (EDA)

A master database of all measurements taken at each sample location in each study area was created. Table 3.1 lists the environmental variables measured at each

sample location, abbreviations used in subsequent sections and the supporting sample size for each attribute measurement. Because of the large number of measurements, generic exploratory graphics were developed using the Splus statistical computing language to summarize and explore the data for useful environmental correlations. The basic univariate, bivariate and multivariate plots are introduced here and illustrated in Section 3.3.

Data Summary

Individual soil attributes (e.g. total carbon, cation exchange capacity, exchangeable sodium percentage) sampled by horizon were summarized using a 2 x 2 matrix of plots. The plots summarize the univariate sample distribution for soil attributes and show bivariate relationships for each attribute by depth of sample and general soil horizon or layer (horizon (x axis) vs. depth (y axis); soil attribute vs. depth; soil attribute vs. probability density; soil attribute vs. horizon). These indicate whether the data are normally distributed, if outliers exist and how the variation is partitioned according to depth and horizon. This enables an initial visual assessment that may indicate the most appropriate modelling approach (e.g. smooth spline by depth, segregated models by horizon). Figures 3.9, 3.15 and 3.22 illustrate these EDA graphics.

Multivariate Exploration and Conditioning

Relationships between response and potential explanatory variables were examined initially using scatterplot matrices. If the data summary plots indicated that horizons usefully partitioned the response variation, scatterplots were developed for subsets of the data based on horizon. Conditioning plots or coplots were developed to simultaneously visualize how variation for individual soil attributes changes both down the profile by depth and horizon and through the landscape as quantified by CTI (e.g. along the environmental gradient). The advantage of these plots is their capacity to reveal complex patterns in multivariate data using environmental gradients. The coplots graph subsets of data in a matrix of panels according to an ordered

Table 3.1 Environmental Variables Measured At Each Sample Location

Environmental Attribute	Abbreviation	Sample Support
Sample Location		
Sample number		
GPS coordinates (AMG easting, northing, elevation)		
Soil Morphological (soil core descriptions)		
Colour		horizon
Hand Texture		horizon
Horizon		core
A horizon depth		core
E horizon presence/absence		core
E horizon depth		core
Mottle presence/absence		core
Depth to mottles		core
Solum depth (A + E + B)		core
Soil material sample depth	DEP	horizon centre
Soil Chemical (horizon depth mid-point soil sample)		
pH		80g
total carbon		80g
exchangeable cations (Ca, Mg, Na, K)		80g
electrical conductivity		80g
Soil Physical (horizon depth mid-point soil sample) (Brucedale and Ladysmith)		
median diameter		80g
model diameter		80g
% gravel		80g
% sand		80g
% silt		80g
% clay		80g
Terrain (computed from 20m DEM)		
elevation	ELEV	20m grid
slope %	SLPP	20m grid
aspect	ASP	20m grid
profile curvature	PRCRV	20m grid
plan curvature	PLCRV	20m grid
tan curvature	TCRV	20m grid
flow accumulation	NCELL	20m grid
cti	CTI	20m grid
flow path length	FPL	20m grid
upslope mean (slope, plan & profile curvature)	MSLP	20m grid
sediment transport index	STRIN	20m grid
stream power index	SPI	20m grid
Airborne Gamma Radiometric Signal (100m and 400m lines interpolated to 50m grid)		
Potassium	K400, K100	50m grid
Thorium	TH400, TH100	50m grid
Uranium	U400, U100	50m grid
Total Count	TC400, TC100	50m grid
Handheld Gamma Radiometer Measurements (Ladysmith)		
K, Th, U, Total Count		~2m radius
Climate (computed from 245m DEM)		
annual mean radiation	AMR	245m grid
precipitation of dryest quarter	PDQ	245m grid
total annual precipitation	TAP	245m grid
maximum temperature warmest month	MAXT	245m grid
mean annual air temperature	MAT	245m grid
minimum temperature coldest month	MINT	245m grid
Digital Orthophoto (scanned and rectified photogrammetry) (Griggward)		
red band	RED	2m grid
green band	GREEN	2m grid
blue band	BLUE	2m grid

classification of a third or fourth variable (e.g. CTI quantile environmental gradient). The classification or conditioning objects are termed "shingles" (Cleveland, 1993) because, like shingles on a roof, the conditioning intervals may overlap.

This enables the elucidation of patterns or, for instance, landscape or catenary thresholds that may fall between two discrete conditioning intervals in attribute space and be missed or obscured due to the intervals chosen for visualization. Shingles may be created based on values in a sample of data, on population parameters (if known) or on arbitrary values chosen through iterative exploration. The study area terrain attribute pdf's provide convenient population models for development of conditioning shingles along terrain environmental gradients in each study area.

Individual shingle sets were created for CTI and its component slope and logarithm of the specific catchment area attributes using the respective population probability density functions for each study area. Values for the cumulative five percent quantiles for each terrain attribute were obtained and shingles with twenty-five and twenty percent overlap established. For example, the Ladysmith CTI pdf (Figure 3.2) was partitioned into overlapping shingles as follows:

```
====      0-20th percentile, 20% of pdf        CTI.pop = 3.77-6.59
    ====      15-40th percentile, 25% of pdf       CTI.pop = 6.44-7.15
        ====      35-60th percentile, 25% of pdf       CTI.pop = 7.00-7.83
            ====      55-80th percentile, 25% of pdf       CTI.pop = 7.64-8.73
                ====      75-100th percentile, 25% of pdf      CTI.pop = 8.46-20.82
```

The shingle sets for CTI, slope and specific catchment area were then used in a function to display soil attribute depth plots. The quantile in terrain attribute space corresponding to each panel of the coplot is indicated by the bar graph above each panel. Plotting symbols were used to indicate the horizon for each sample point and a line connecting respective points for each profile added to provide perception of profile sample connectivity. The overall visual pattern of the data through the panels of the plot matrix convey a gestalt understanding of soil attribute variation down hillslopes or through the environmental attribute spaces representative of the study area.

Figures 3.10, 3.16, 3,17, 3.23 and 3.24 illustrate coplots. Note the dark bar at the head of each panel that indicates the range of each shingle.

Exploratory Trees

Exploratory Trees were created for individual response soil attributes with a range of explanatory environmental attributes using Splus (Statistical Sciences, 1993). The Splus Tree function uses a recursive binary partioning algorithm that chooses the best set of "splits" that partition the response variable space into increasingly homogeneous sets. The vertical length of the Tree branches is an indication of the reduction in deviance obtained by each node of the Tree. The final branch splits often reduce deviance by only small amounts, suggesting that the these conditional relationships are not as important and should be interpreted cautiously. Conditional rules can be simply derived from the Tree plots and these provide predictions of the response variable. The fitted model in conjunction with map algebra tools may then be used to create spatial displays of the predicted variable. Figures 3.5, 3.11 and 3.26 illustrate exploratory Trees.

3.2.7 Statistical Modelling

Modelling Criteria

The intention of the above EDA is to thoroughly evaluate univariate, bivariate and multivariate environmental correlations that may be used for prediction. This was followed by development of statistical models using the following iterative series of steps:

- implement a stepwise explanatory attribute selection algorithm using the AIC (Akaike Information Criterion);

- develop a model based on explicit decisions from previous EDA or stepwise selection process;

- evaluate the model with diagnostics (residual plots, % reduction in residual deviance, degrees of freedom consumed, evaluate tradeoffs between simplicity, complexity and landscape process interpretations); and

- create spatial display.

If the exploratory plots elucidated simple or smooth relationships, model terms were developed (e.g. linear fit, scatterplot smoother) to incorporate them. Tree based models were used to detect non-linear or conditional relationships in the data and as an indicator of the most useful explanatory variables. The Splus automated stepwise attribute selection algorithm (step.gam) was used for selection of potential explanatory attributes to further guide model development. The algorithm ranks individual explanatory attribute terms and all possible combinations of input terms by an AIC statistic that effectively balances the reduction in residual deviance by the degrees of freedom consumed.

The percentage reduction in residual deviance is a quantitative indicator of the proportion of deviance or variance accounted for and suggestive of the goodness of fit or level of certainty of a model (McCullagh and Nelder, 1989). It represents a percentage improvement from a null model that would simply use the mean value for a variable. Therefore, it suggests the improvement over taking the mean value for a variable using the collected sample points within the bedrock geology map unit. In cases where the response variable is continuous and the error model is normal, percentage reduction in residual deviance (%RID) is equivalent to the multivariate R^2 used widely in traditional statistics.

The step.gam function allows a broad range of parametric and non-parametric term fits to be evaluated (e.g. mean, linear, smoothing spline, loess, natural cubic splines, polynomial splines). Each method used for fitting a term also has several smoothing parameters that may be adjusted to provide better fits. Because of the large number of models being developed for this work, the step.gam function was systematically implemented incorporating explanatory variables as linear, smoothing spline and loess (local regression) fit terms using default parameter settings.

In the end, judgements were required to decide which models to use and how to interpret relationships discovered. These will be discussed below.

Spatial Implementation

Digital maps were created using developed models when possible. Development of a spatial implementation required that:

- all explanatory variables be available over the study area in a spatially continuous manner; and
- the model terms or coefficients be expressed in a manner that could be implemented using map algebra.

Both GLM and Tree based models provided coefficient terms (GLM) and conditional branching relationships that could be simply implemented in the GIS. GAM's using non-parametric terms such as scatterplot smooothers required extra steps to develop prediction lookup tables using all possible combinations of explanatory variable values. These were much more laborious to generate, but feasible as demonstrated in the profile total carbon model below. The spatial implementations may then be draped over DEM's for visual assessment.

3.3 RESULTS AND DISCUSSION

A subset of soil variables (solum depth, total carbon, cation exchange capacity and exchangeable sodium percentage) from the Ladysmith study area is used to illustrate the model building approach and test the chapter hypothesis. Of the eighty-five sample sites selected in Ladysmith, seventy three locations were sampled and twelve locations ruled out due to logistic restrictions (e.g. roads, buildings, denied access).

3.3.1 Solum Depth

Solum depth is defined here as the depth of the A plus B horizon(s) expressed in centimetres below the surface (negative). In general it provides an indication of water storage capacity, nutrient pools, overall productivity and has management implications for plow and plant rooting depths, erosion and other land degradation

processes. It is one of the critical soil layer attributes used to develop integrated landscape models of other soil attributes.

Exploratory Plots

Univariate probability density and Q-Q plots of solum depth showed a multimodal distribution with a broad cluster of samples around the 50cm depth and a tighter cluster around the 200cm depth. Pairwise scatterplots of solum depth versus terrain, radiometric and climatic explanatory variables indicated useful relationships with a subset of primary and secondary terrain attributes and radiometric potassium. Figure 3.4 shows these relationships via scatterplots with a linear least squares model fit (hashed line), a scatterplot smoother using a loess locally quadratic fitting method (solid line) and a scatterplot smoother using a smooth spline (dotted line). Solum depth fitted linearly by CTI consumes two degrees of freedom with a %RID of 58. Solum depth fitted by the loess model consumes 5.4 degrees of freedom and provides a %RID of 75, while the smooth spline consumes five degrees of freedom with a %RID of 74. These indicate that CTI alone accounts for a large proportion of the variation in the solum depth sample set. Equivalent comparisons may be done for each bivariate relationship.

Figure 3.5 illustrates a regression Tree predicting solum depth using all available environmental variables (Table 3.1). The terminal nodes or leaves produce predictions of solum depth in centimetres. If the total reduction in residual deviance is judged similarly to the GAM models, this model achieves a %RID of 90 consuming five degrees of freedom producing nine terminal leaves. This indicates the binary splits achieve a high level of homogeneity at the terminal nodes. The branch length to terminal leaves for the deeper soils indicates they are the least homogeneous (least predictable) or within group variability is the highest. The mid-depth soils are slightly more homogeneous than the shallow soils. The airborne K400 radiometric signal provides additional discrimination of solum depth at the mid-depths. This provides a useful complement to the terrain attributes, primarily CTI. The Tree does

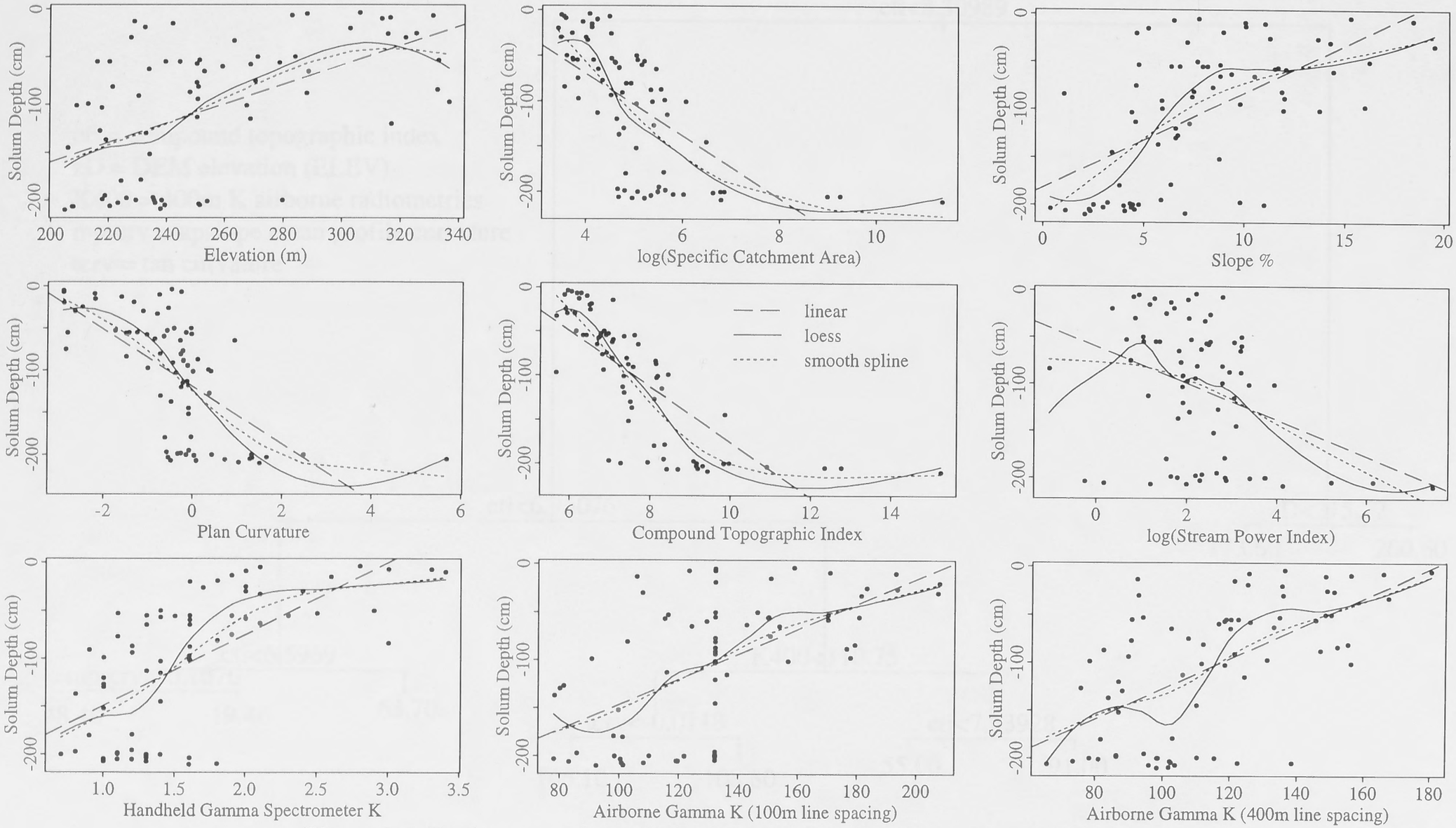


Figure 3.4 Environmental Attributes (Explanatory) vs. Solum Depth (Response)

cti = compound topographic index
zD = DEM elevation (ELEV)
K400 = 400m K airborne radiometrics
mprcrv = upslope mean profile curvature
tcrv = tan curvature

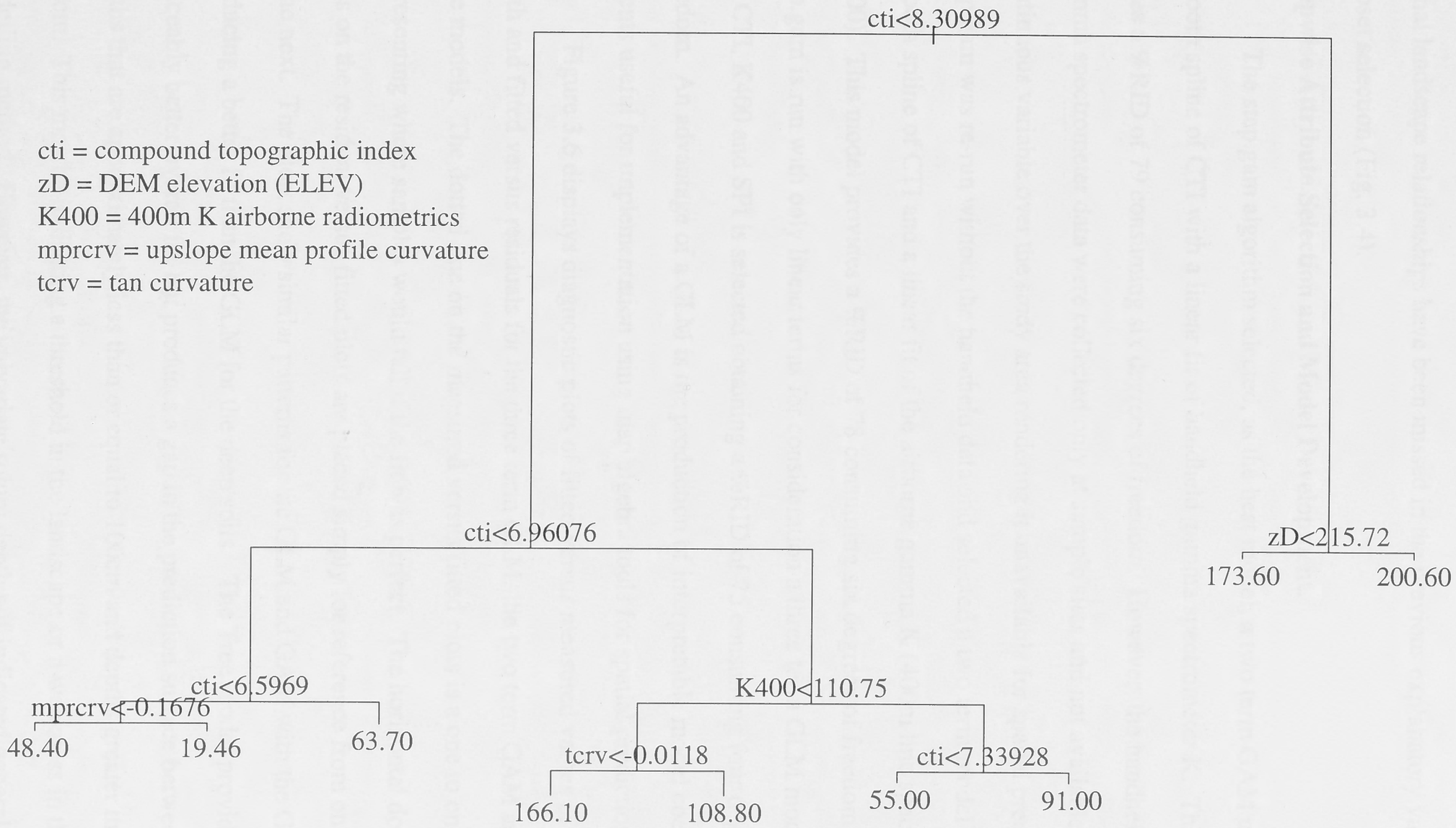


Figure 3.5 Solum Depth Regression Tree Model

not suggest that important explanatory variables indicative of non-linear or conditional landscape relationships have been missed in the previous explanatory variable subset selection (Fig. 3.4).

Stepwise Attribute Selection and Model Development.

The step.gam algorithm selected, as the best model, a two term GAM using a smooth spline of CTI with a linear fit of handheld gamma spectrometer K. This provides a %RID of 79 consuming six degrees of freedom. However, the handheld gamma spectrometer data were collected only at sample sites and not available as a continuous variable over the study area rendering it unavailable for spatial prediction. Step.gam was re-run without the handheld data and selected a two term model with a smooth spline of CTI and a linear fit of the airborne gamma K (400m line spacing - K400). This model provides a %RID of 78 consuming six degrees of freedom. If step.gam is run with only linear terms for consideration a three term GLM model using CTI, K400 and SPI is selected obtaining a %RID of 75 consuming four degrees of freedom. An advantage of a GLM is the production of interpretable model coefficients useful for implementation using map algebra tools for spatial prediction.

Figure 3.6 displays diagnostic plots of fitted versus measured values of solum depth and fitted versus residuals for the three term GLM, the two term GAM and the Tree models. The dotted line on the measured versus fitted plots is a one to one line representing where samples would fall if the fit was perfect. The horizontal dotted lines on the residual versus fitted plots are placed simply for reference from one plot to the next. The plots show similar patterns for the GLM and GAM with the GAM producing a better fit than the GLM for the deep soils. The Tree model provides a noticeably better overall fit, but produces a gap in the prediction surface between depths that are approximately less than or equal to 100cm and depths greater than 160cm. This may be indicating a threshold in the landscape or a weakness in the modelling method. However, the univariate solum depth pdf indicated general clusters around 50cm and 200cm. This may suggest two process related solum depth

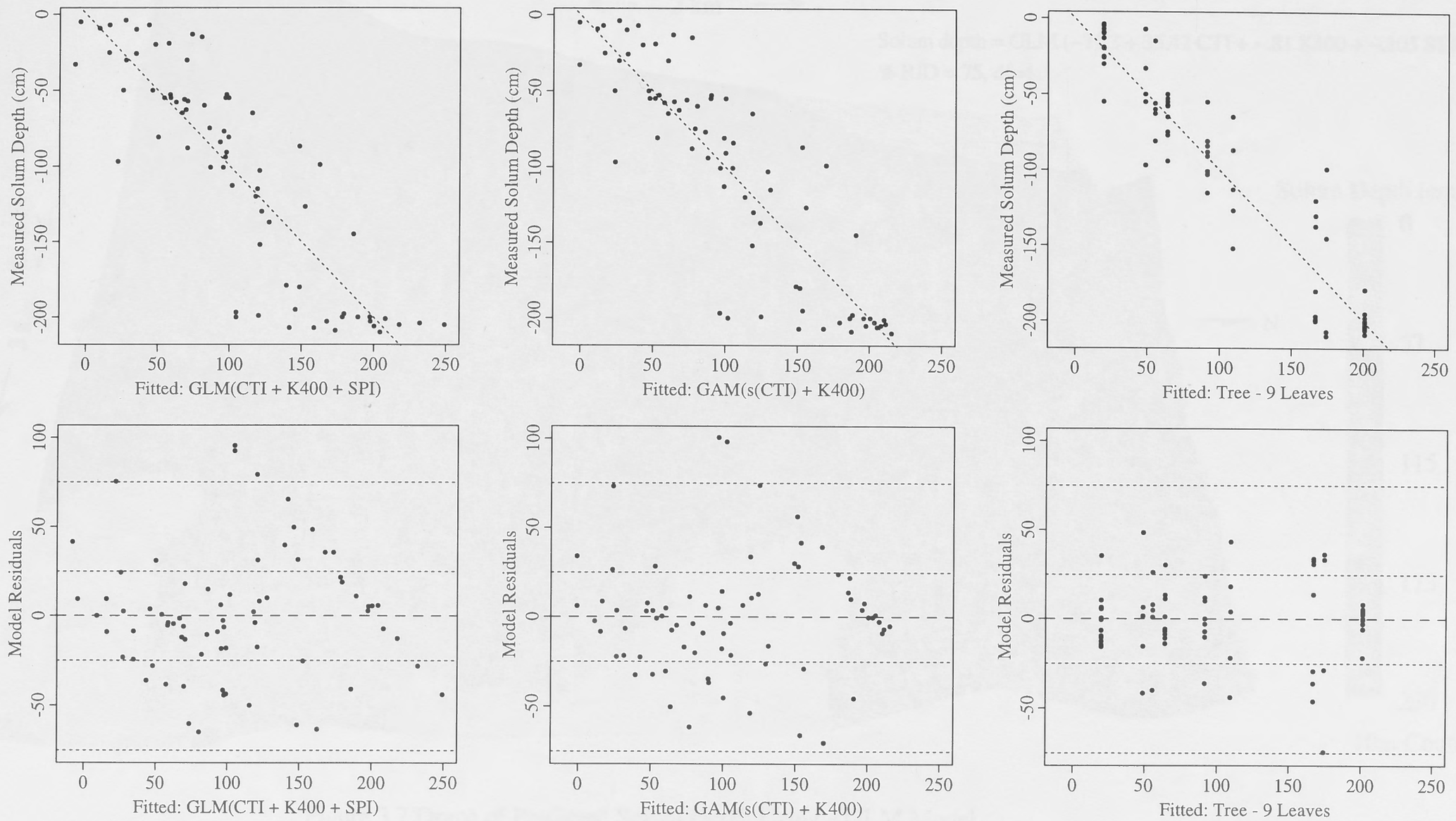


Figure 3.6 Fitted Solum Depth vs. Measured and Fitted Solum Depth vs. Residuals

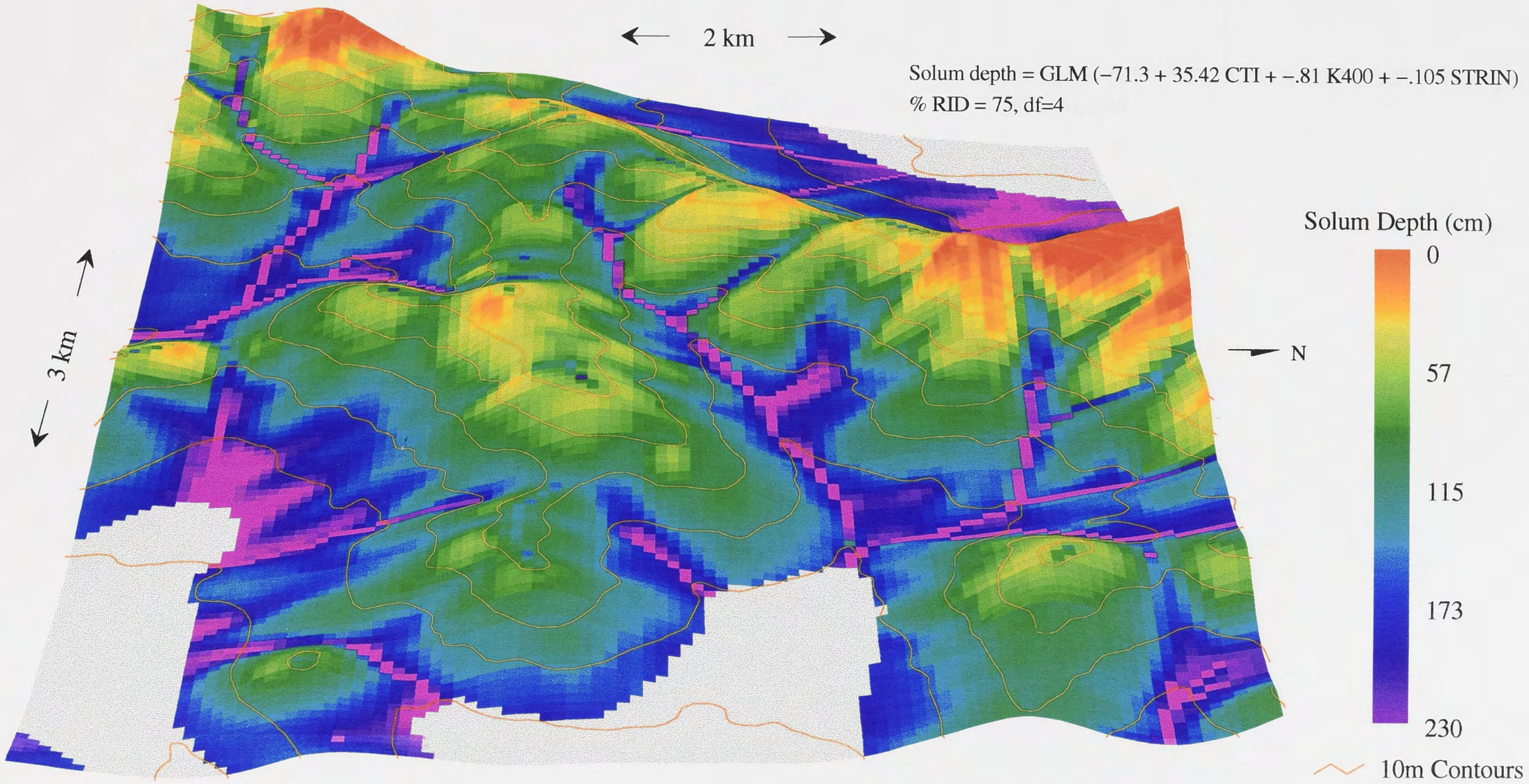


Figure 3.7 Drape of Predicted Solum Depth Using a GLM Model

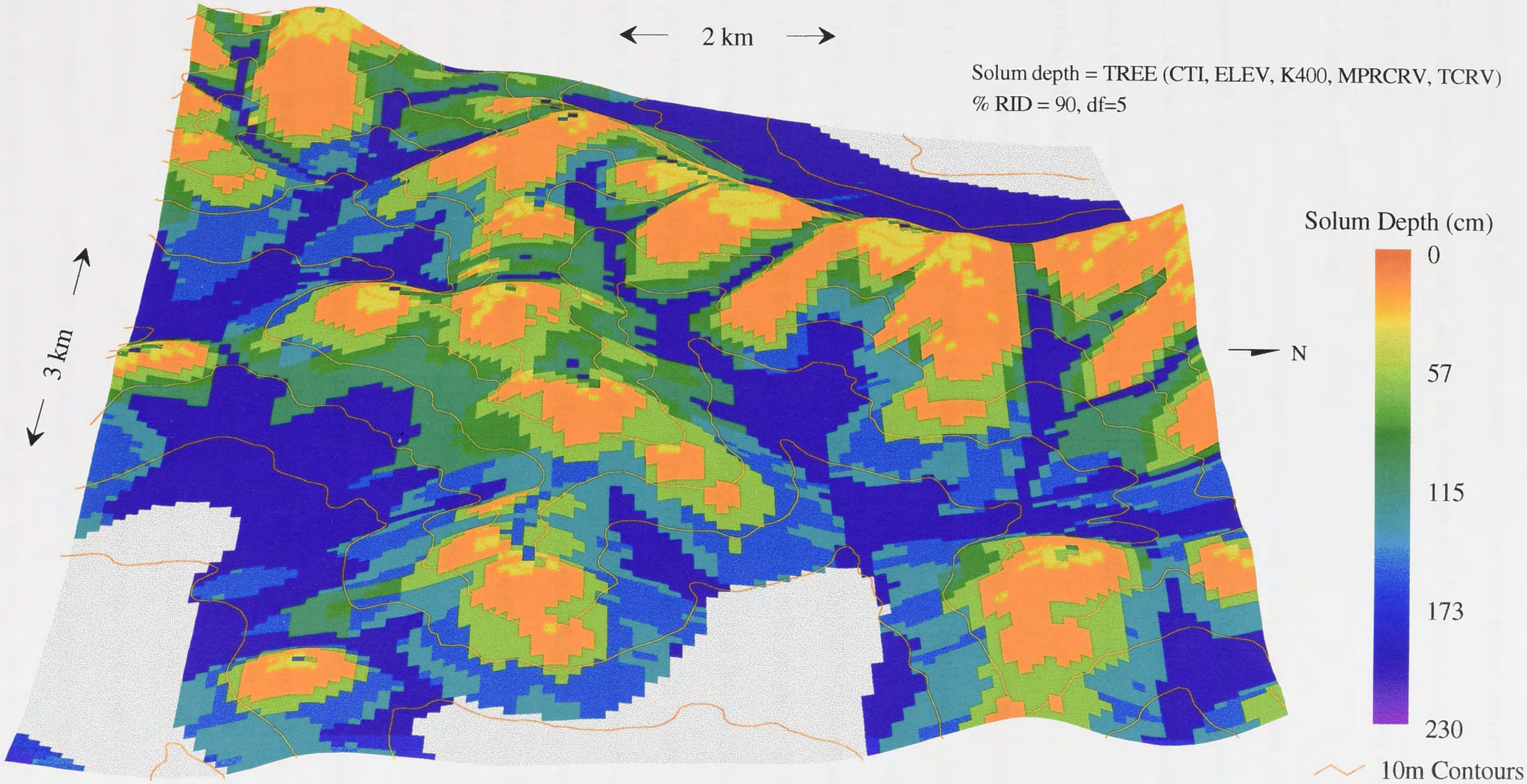


Figure 3.8 Drape of Predicted Solum Depth Using a TREE Model

sub-populations relating to shallower, erosional soils in the upper part of the landscape and deeper, depositional soils in the lower parts of the landscape. The middle of the landscape may be a transferal zone where the gamma radiometric data play a more useful predictive role.

The plot of the Tree residuals provides a visual indication of the within group variance at each of the nine leaves. Both the GLM and GAM methods produce two positive solum depth values with the GAM values closer to zero. The plots do not indicate that a change from the normal error model is warranted.

Spatial Display

Figures 3.7 and 3.8 illustrate the spatial implementation of the three term GLM model and the five explanatory variable Tree model draped over a DEM with 10m surface elevation contours indicating relief. The stepped prediction surface of the Tree is apparent with comparison indicating that the Tree does not model solum depth extremes (e.g. shallow soils on ridge tops and deep soils in valley bottoms) in the same manner as the GLM. The Tree constricts the range of the predicted solum depth attribute space or produces a smoother prediction surface. In general, the Tree appears to predict larger areas of shallow and deep soils, which match the sample pdf better than the GLM implementation. However, the overall %RID's for each of the three models suggest that solum depth can be predicted with a high level of certainty and low level of complexity using any of these models.

3.3.2 Total Carbon

Total carbon is expressed in mass percent (g/100g soil) and is directly related to organic matter content when carbonates are not present. Moderate levels of organic matter are required for the maintenance of soil structure and total carbon is an overall indicator of biological activity and soil health. High levels improve cation exchange capacity, pH buffer capacity and soil pollutant attenuation and complexation. Better estimates of total carbon are required as inputs into simulation models for a range of purposes (e.g. crop production, carbon cycling etc.). The soil has great

potential for sequestration of atmospheric carbon, but current methods for predicting distribution are inadequate.

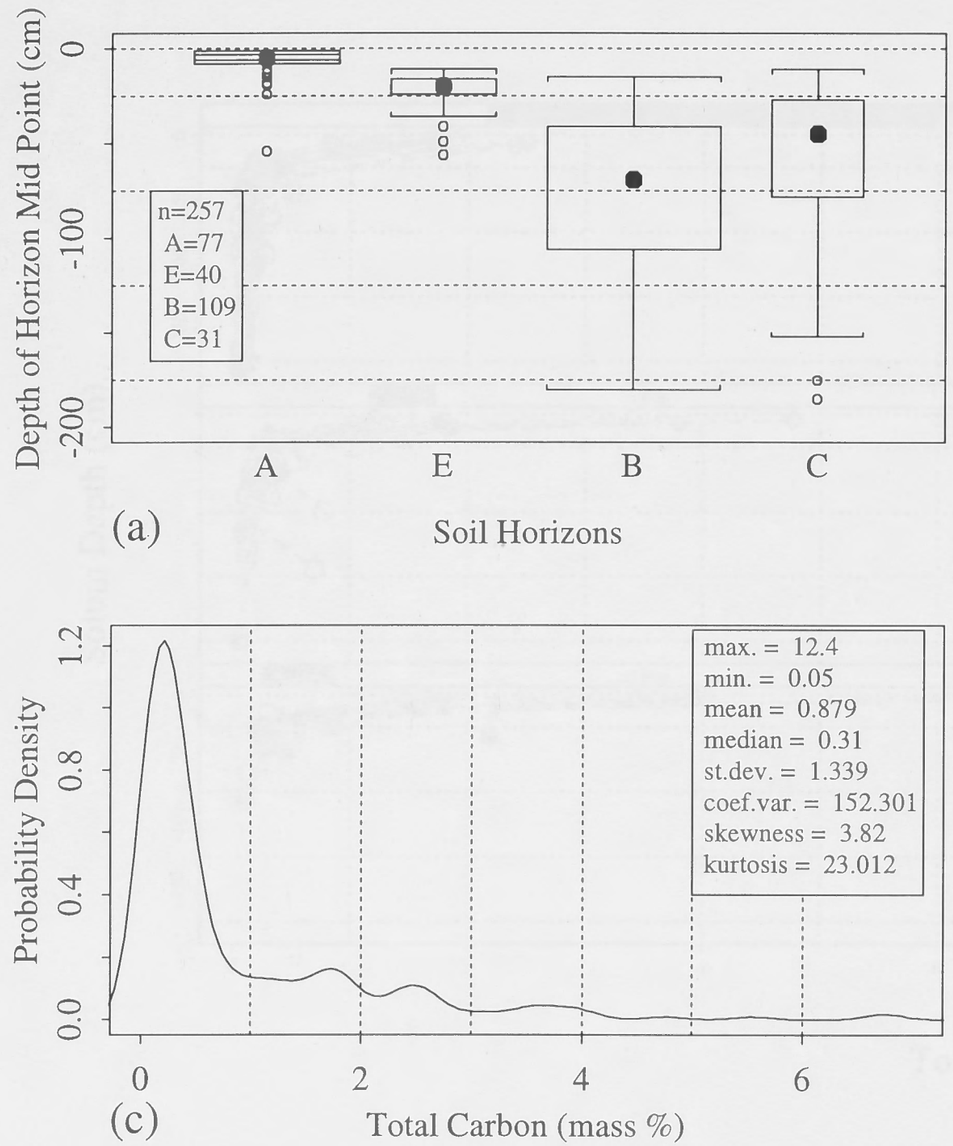
Exploratory Plots

Figure 3.9 shows univariate and bivariate exploratory data analysis plots illustrating total carbon relationships with sample depth and soil horizon. An outlier A horizon sample with a total carbon mass percent of 12.4 is outside the bounds of the EDA plots. The sample distribution (Fig. 3.9c) is strongly peaked and positively skewed. Although a few outliers exist, Figure 3.9b indicates that total carbon exhibits a smooth relationship with depth suggesting that a scatterplot smooth GAM model may be an appropriate approach. Figure 3.9d shows that most of the carbon for the Ladysmith landscape is stored in the A horizons. While the soil horizons are statistically significant in partitioning the variation as visualized by Fig. 3.9d, it is more efficient to use depth as a predictor due to the smooth depth relationship across horizon bounds.

Figure 3.10 shows a coplot of total carbon conditioned by CTI shingles (as discussed above) providing an illustration of the total carbon soil profiles down the hillslope continuum. The panels are organized by CTI conditioning interval from left to right and bottom to top according to high landscape positions (small CTI - lower left panel) to low landscape positions (large CTI - upper left panel) as indicated by the shingle strips (labelled cti.pop) above each panel. The dashed lines connect sample points from individual soil cores. Although A horizon total carbon appears to decrease in mean and variance down the toposequence, the generally smooth total carbon relationship with depth is invariant with landscape position.

Stepwise Attribute Selection and Model Development

Based on relationships indicated above, two predictive models were developed. One for prediction of A horizon total carbon, where most of the carbon is located, and a second for prediction of total carbon held in the soil profile using an



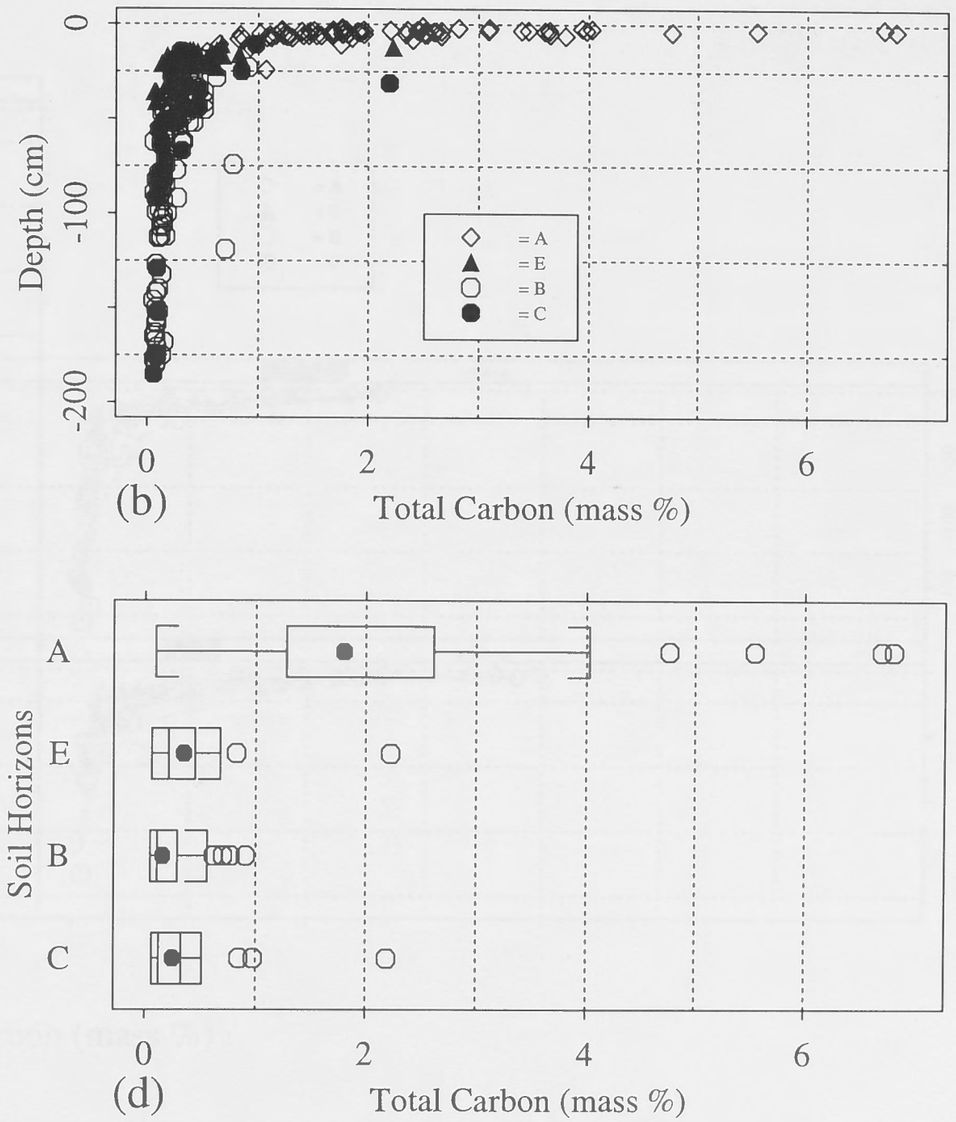


Figure 3.9 (a-d) Total Carbon Univariate and Bivariate EDA (Ladysmith)

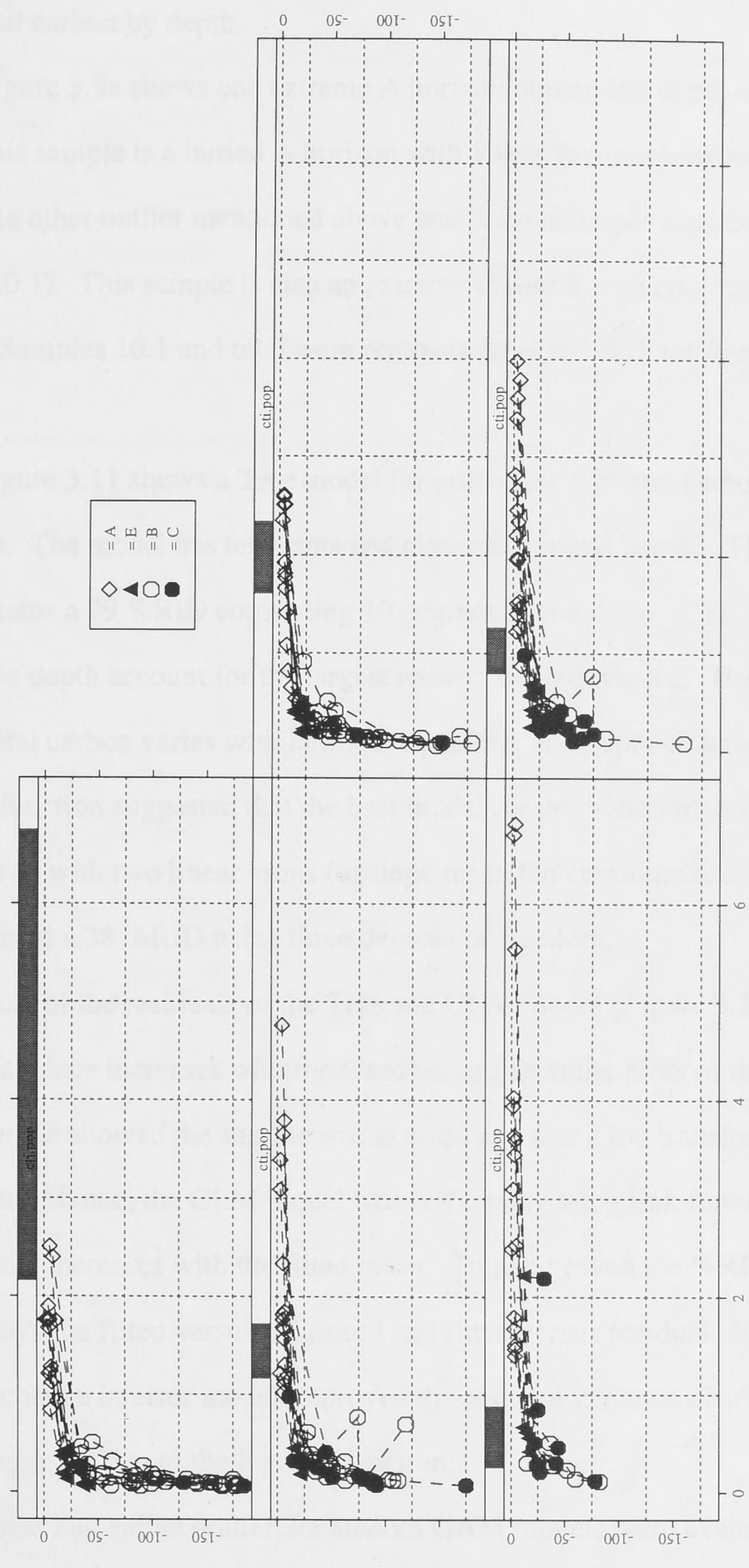


Figure 3.10 Total Carbon Coplot Conditioned by Compound Topographic Index (Ladysmith)

integration of A horizon and solum depth models with a scatterplot smoother GAM fitting total carbon by depth.

Figure 3.9a shows one extreme A horizon outlier at a depth of 54cm (sample 69.3). This sample is a buried A horizon with a very low total carbon content of 0.1%. The other outlier mentioned above was from a sample containing charcoal (sample 10.1). This sample is also an extreme outlier for several other chemical attributes. Samples 10.1 and 69.3 were removed from the data set prior to model development.

Figure 3.11 shows a Tree model for predicting the total carbon content of the A horizon. The model has ten terms and thirteen terminal leaves. This complex model obtains a 79 %RID consuming 10 degrees of freedom. CTI, flow path length and sample depth account for the largest reductions in deviance. This indicates that A horizon total carbon varies with landscape position and depth of sample. The step.gam function suggested that the best model for prediction of A horizon total carbon was a fit with two linear terms (upslope mean tan curvature and sample depth). This provided a 38 %RID using three degrees of freedom.

Plots of the residuals of the Tree and GLM model (Figure 3.12) indicate that residual deviance increases with the fitted mean. Quantile plots of the A horizon total carbon sample showed the sample was skewed and that a log transformation may be appropriate. Hence, the GLM model was re-fit using a log link function with an error variance that increases with the fitted mean. This improved the %RID to 48. Figure 3.12 displays the fitted versus measured and fitted versus residuals for the three models. The change in error model improved the residual variance distribution with one outlier magnified due to the log link function.

Loess and spline scatterplot smooth GAM models were evaluated for predicting the solum total carbon profile. Figure 3.13 illustrates the scatterplot smooths (loess - top row, spline - bottom row) and fitted versus both measured values and residuals for each model. The left plots show scatterplots of the sample points with the

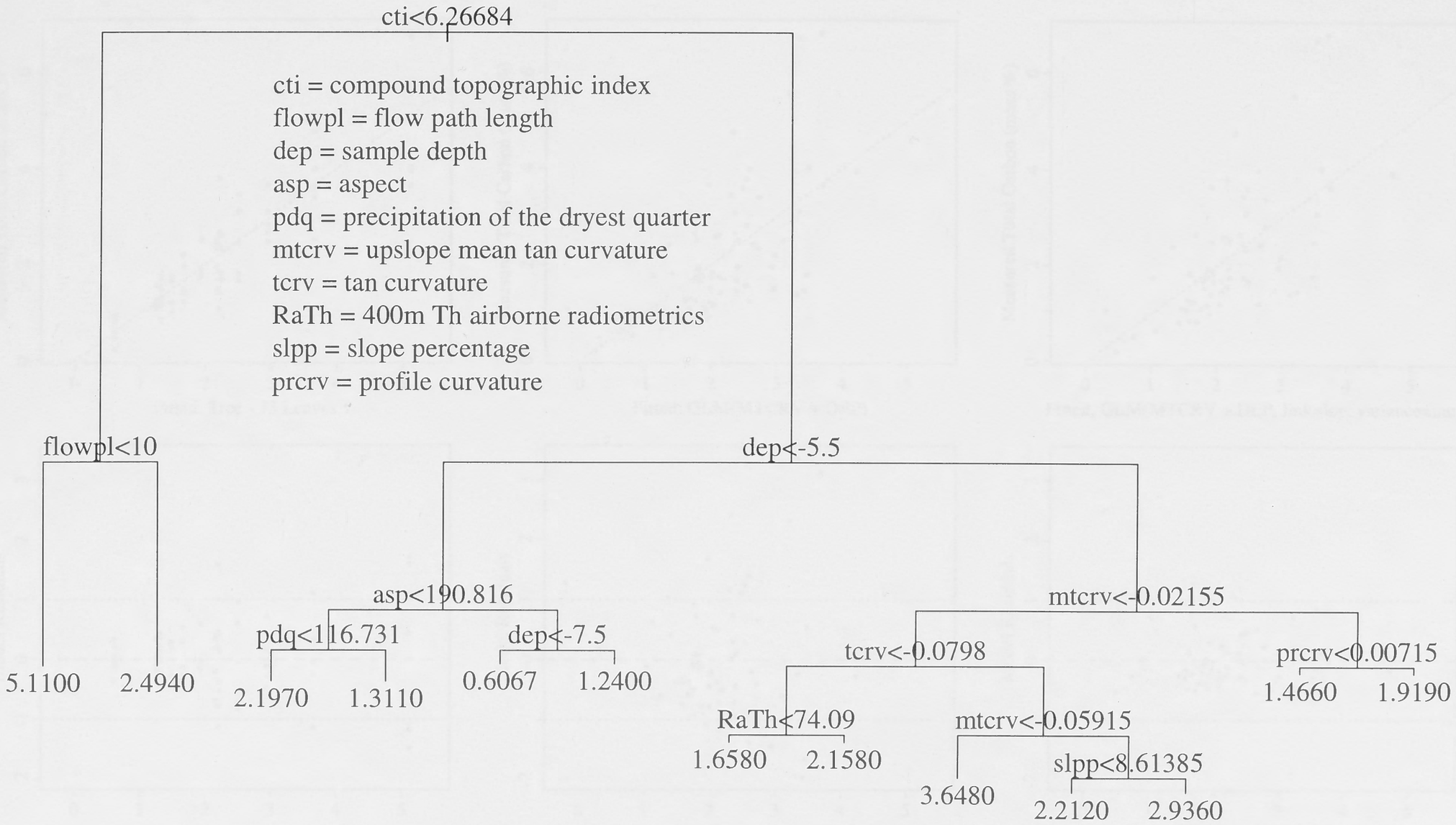


Figure 3.11 A Horizon Total Carbon Regression Tree Model

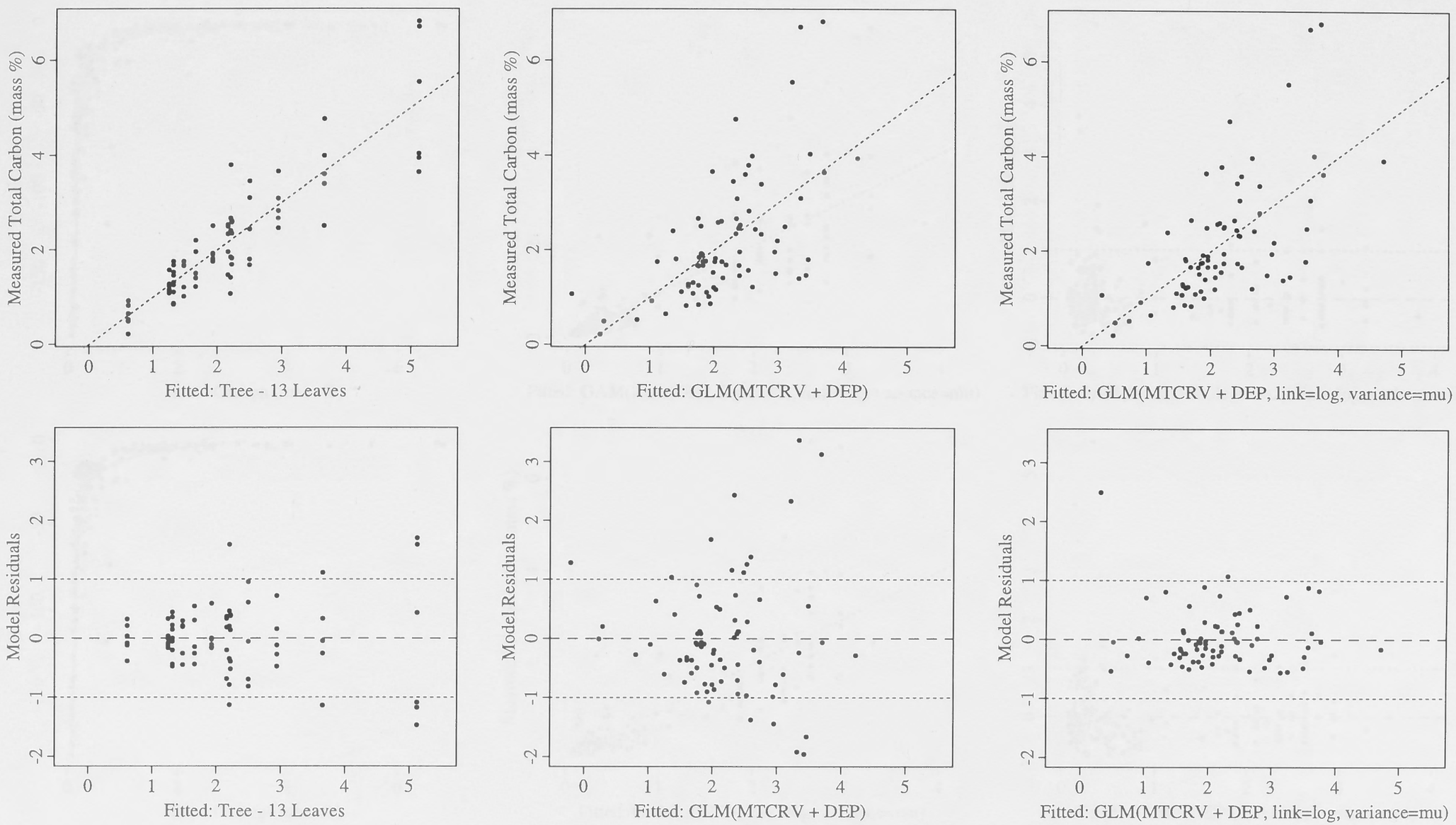


Figure 3.12 Fitted A Horizon Total Carbon vs. Measured and Fitted A Horizon Total Carbon vs. Residuals

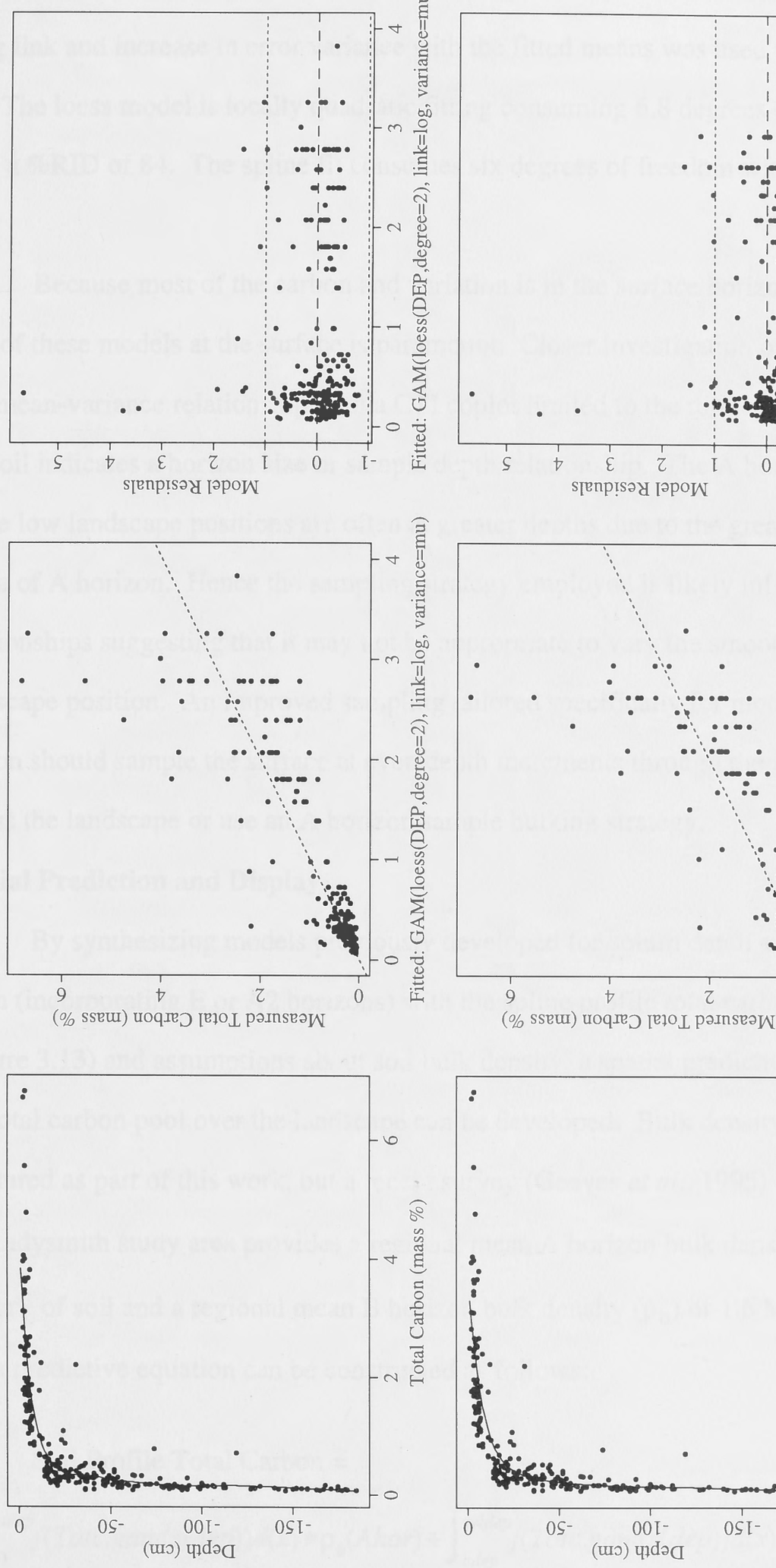


Figure 3.13 Fitted Profile Total Carbon vs. Measured and Fitted Profile Total Carbon vs. Residuals

fitted models (solid line). The residual variance increases with the fitted mean, and so a log link and increase in error variance with the fitted means was used for both models. The loess model is locally quadratic fitting consuming 6.8 degrees of freedom with a %RID of 84. The spline fit consumes six degrees of freedom with a %RID of 83.

Because most of the carbon and variation is in the surface horizons, the behaviour of these models at the surface is paramount. Closer investigation of the A horizon mean-variance relationship with a CTI coplot limited to the top ten centimeters of the soil indicates a horizon size or sample depth relationship. The A horizon samples in the low landscape positions are often at greater depths due to the greater overall depth of A horizon. Hence the sampling strategy employed is likely influencing these relationships suggesting that it may not be appropriate to vary the smooth fit with landscape position. An improved sampling tailored specifically for modelling total carbon should sample the surface at even depth increments through the A horizon across the landscape or use an A horizon sample bulking strategy.

Spatial Prediction and Display

By synthesizing models previously developed for solum depth and A horizon depth (incorporating E or A2 horizons) with the spline profile total carbon model (Figure 3.13) and assumptions about soil bulk density, a spatial prediction of the profile total carbon pool over the landscape can be developed. Bulk density was not measured as part of this work, but a recent survey (Geeves *et al.*, 1995) encompassing the Ladysmith study area provides a regional mean A horizon bulk density (ρ_b) of 1.5 Mg m^{-3} of soil and a regional mean B horizon bulk density (ρ_b) of 1.6 Mg m^{-3}. With this a predictive equation can be constructed as follows:

Soil Profile Total Carbon =

$$\int_0^{adep} f(Totc.gam(s(dep))d(x) * \rho_b(Ahor) + \int_{adep}^{soldep} f(Totc.gam(s(dep))d(x) * \rho_b(Bhor)$$

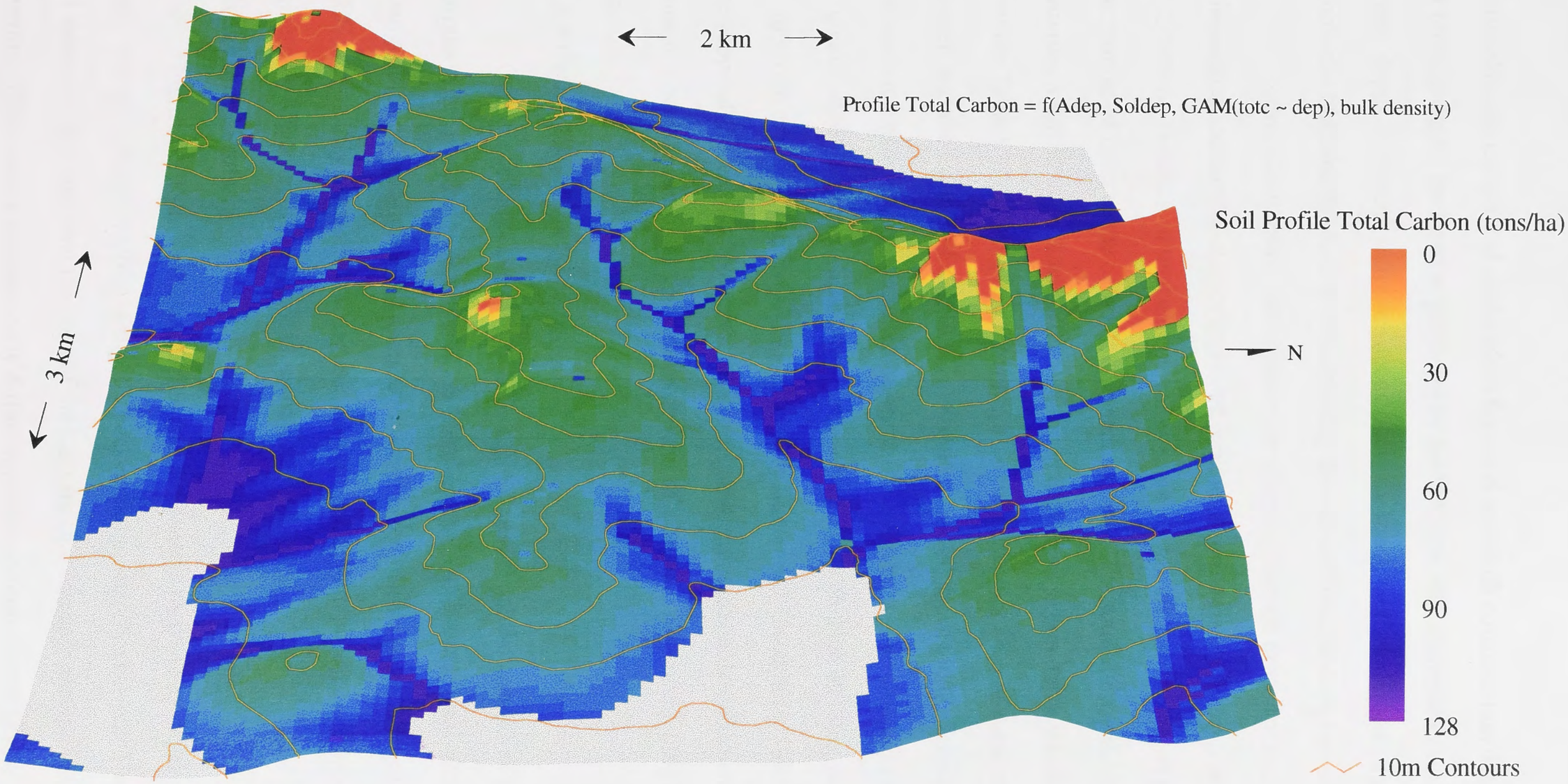


Figure 3.14 Drape of Predicted Soil Profile Total Carbon

This takes the integrals of total carbon for the A and B (B equals solum depth minus A horizon depth) horizons as predicted by the GAM model, computed on a one centimeter depth increment, times the bulk density for the A and B horizons. This equation was computed for each grid node using the values from the component models.

Figure 3.14 displays the predicted profile total carbon draped across a DEM. The display indicates that once soil depth reaches approximately 30cm (see Figures 3.7, 3.8) the variation in total carbon varies more slowly with solum depth reflecting the storage of most of the carbon in the near surface. The total carbon GAM model provides a %RID of 84% while the A horizon depth and solum depth GLM models provide 78 and 77 %RID's respectively. This indicates that each of these component models and the integrated quantitative model of the soil profile total carbon pool can be predicted with a high level of certainty in this landscape. Development of methods to quantitatively represent how error is propagated through such an integrated model is an active area of research (Heuvelink, 1993; Hunter and Goodchild, 1995) beyond the scope of this thesis. However, the explicit and quantitative development of component models is the first step to development of a broader error modelling methodology.

3.3.3 Cation Exchange Capacity

Cation exchange capacity (CEC) is an expression of the number of cation adsorption sites per unit weight of soil. It is defined here as the sum total of basic exchangeable cations adsorbed, expressed in centimoles of charge per kilogram of soil. CEC is primarily controlled by organic matter content and quantity and type of clay minerals. It is a useful indicator of soil chemical fertility.

Exploratory Plots

Figure 3.15 shows the univariate and bivariate EDA plots by depth and horizon. The sample distribution is slightly peaked and positively skewed. The range (1-21 cmol/kg) indicates that the soils of the Ladysmith study area are low in chemical fertility. CEC shows a distinctively different pattern to total carbon (Fig. 3.9) by not

exhibiting a smooth relationship with depth, but is broadly scattered with E horizons occupying the most tightly clustered area of depth and CEC attribute space. Figure 3.15d illustrates that the horizons (excluding C) provide a statistically significant partitioning of the CEC variation. Hence, a scatterplot smoother, as for total carbon, is not feasible and modelling stratified CEC subsets by horizon is likely the most appropriate approach. The boxplots of the A and E horizon CEC (Figure 3.15d) have a slight positive skew, indicating that a log link may be useful, while the B horizon distribution appears nearly normal and widely scattered (high variance).

Figure 3.16 shows a coplot of CEC conditioned by CTI. The CEC variation does not exhibit apparent hillslope thresholds, but changes gradually as deeper B horizons occur in the landscape. Figure 3.17 displays the CEC profiles as conditioned by both slope and specific catchment area, the components of CTI. These plots systematically condition by slope shingles increasing from left to right and specific catchment area shingles increasing from bottom to top. Again, no distinctive patterns emerge to suggest that the processes influencing CEC exhibit sharp breaks or landscape thresholds in this study area. B horizon CEC's do increase slightly in larger specific catchment area panels, perhaps indicating clay translocation in the landscape or greater *in situ* synthesis at more moist sites. The result of the EDA suggests that CEC is best modelled by individual horizons and, without strong apparent landscape relationships, greater reliance must be placed on automated explanatory variable selection (step.gam).

Stepwise Attribute Selection and Model Development

Separate models for the A, E and B horizons were developed. For A horizon CEC prediction, the step.gam function identified a three linear term fit with upslope mean tan curvature, flow accumulation and U400. This model provided a %RID of 33 consuming four degrees of freedom. An A horizon CEC Tree model produced a %RID of 74 consuming seven degrees of freedom with nine terminal leaves. Theseven attributes used for binary paritioning, in order of largest to smallest reductions

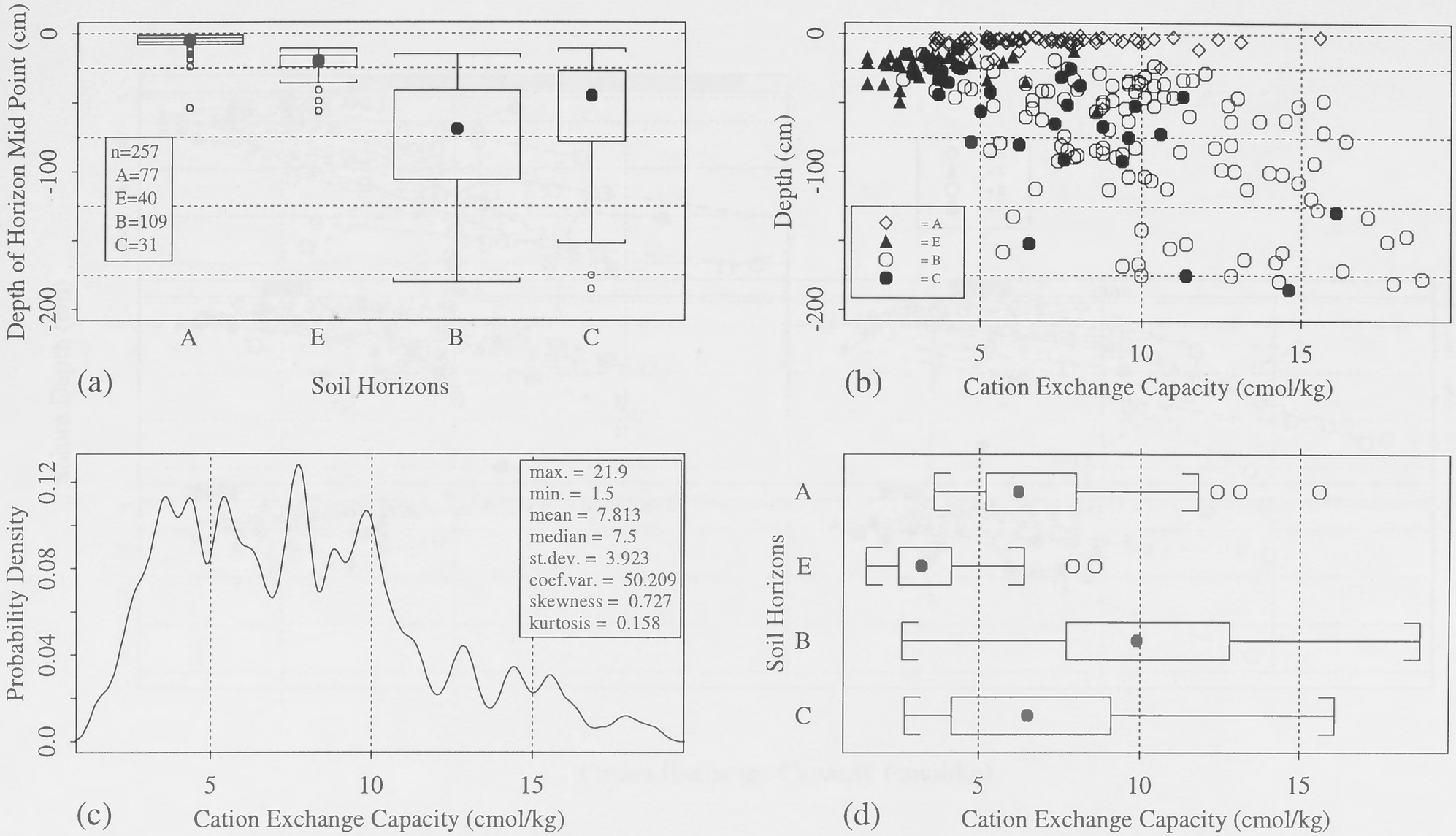


Figure 3.15 (a-d) Cation Exchange Capacity Univariate and Bivariate EDA (Ladysmith)

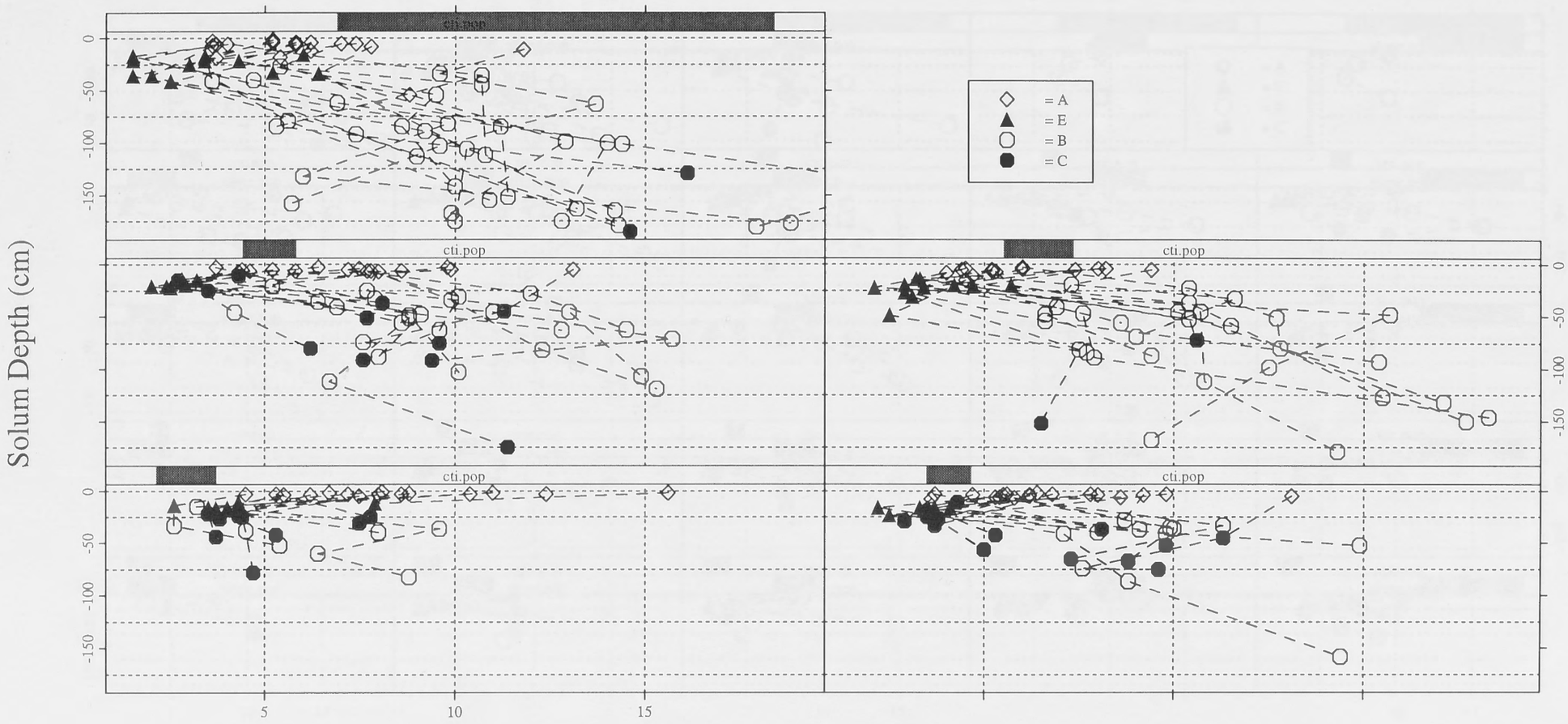


Figure 3.16 Cation Exchange Capacity Coplot Conditioned by Compound Topographic Index (Ladysmith)

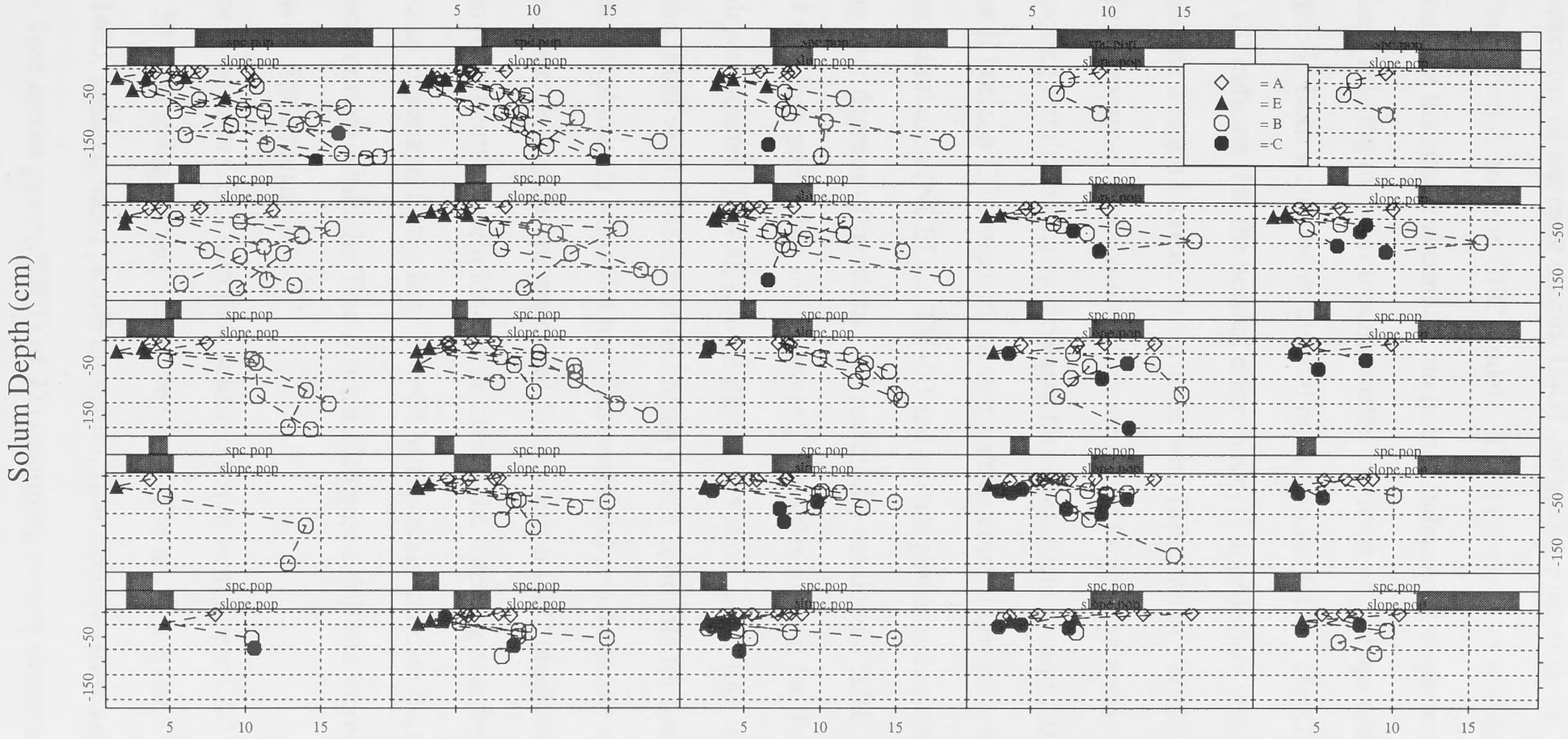


Figure 3.17 Cation Exchange Capacity Coplot Conditioned by Slope and Specific Catchment Area

in deviance were: CTI, tan curvature, upslope mean profile curvature, TH100, maximum temperature of warmest month, U400 and upslope mean slope. Figure 3.18 illustrates these two A horizon CEC models. The explanatory variables selected vaguely suggest that water accumulation in the landscape is important to A horizon CEC perhaps relating to either organic matter contents or perhaps erosion and deposition of topsoil. Both would be influential on A horizon CEC. The GLM residuals (Figure 3.18) indicate that the normal error model is appropriate while the Tree residuals indicate a slight increase in variance with fitted mean. While a log transformation of the response A horizon CEC controls the large Tree residuals when CEC's are 5.0 cmol/kg but dramatically increases those residuals around 2.0 cmol/kg.

For E horizon CEC prediction, step.gam selected a three linear term GLM using flow accumulation (NCELL), TH400 and stream power index with a %RID of 41 consuming four degrees of freedom. A Tree model obtained a %RID of 45 using five degrees of freedom producing seven leaves. Variables used in the Tree were: precipitation of dryest quarter, TC400, sample depth, elevation and U400. Figure 3.19 illustrates the fitted versus measured and fitted versus residuals for these two models. The selected explanatory variables hint that water and geochemistry as reflected in gamma radiometric signals are important for E horizon CEC prediction. This suggests an environmental correlation that relates to the leaching and depletion processes that influence E horizon formation. The GLM residuals indicate the normal error model is appropriate while the Tree residuals suggest a definite increase of residual variance with fitted mean.

For B horizon CEC, step.gam selected a five term linear GLM that obtains a %RID of 39 consuming six degrees of freedom. The attributes used were sample depth, TH400, upslope mean profile curvature, flow accumulation and flow path length. A Tree model gave a %RID of 72 consuming twelve degrees of freedom and produced eighteen leaves. Variables used in order of deviance reduction were:

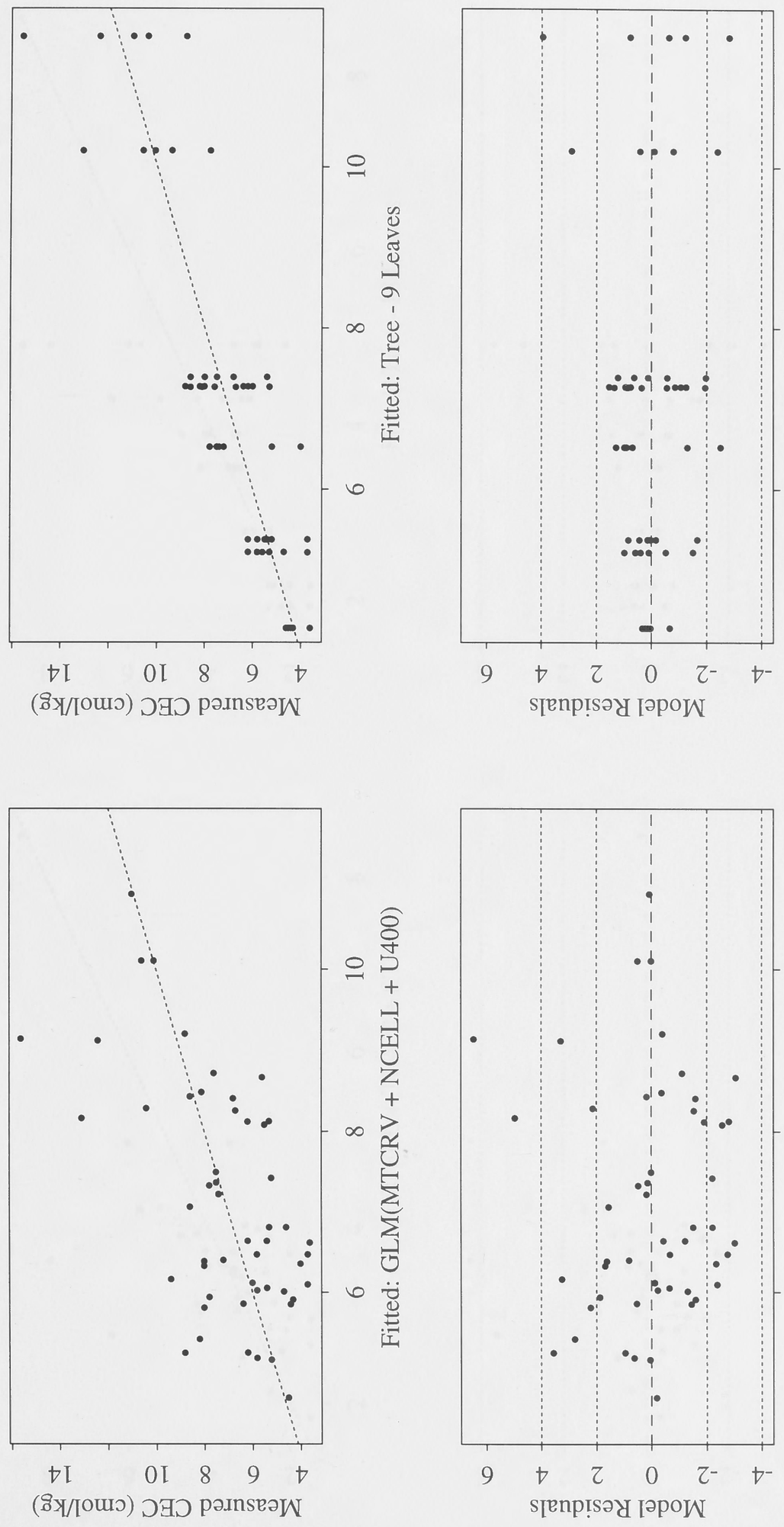


Figure 3.18 Fitted A Horizon CEC vs. Measured and Fitted A Horizon CEC vs. Residuals

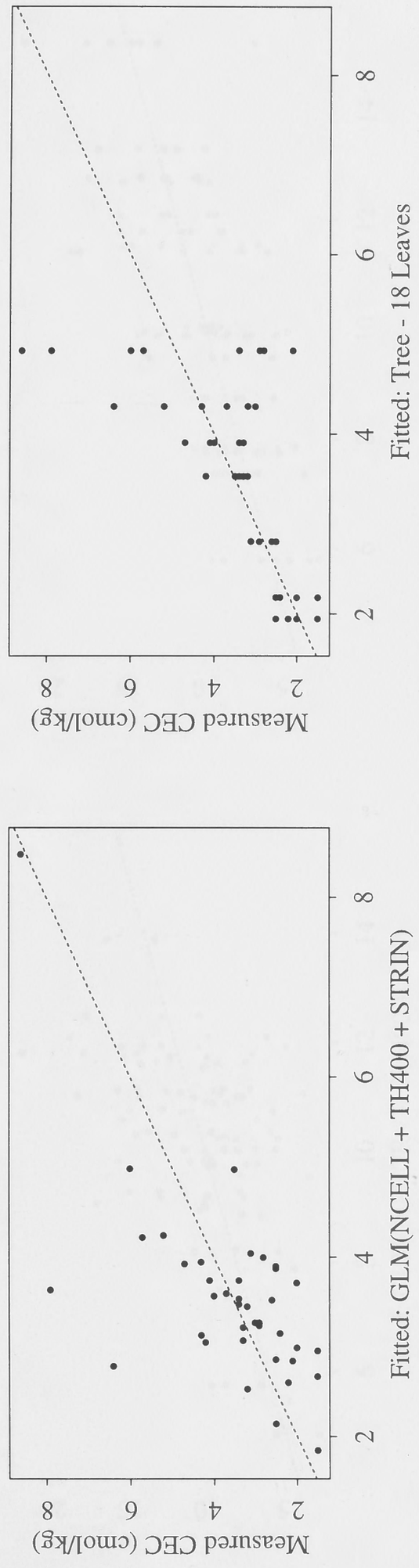


Figure 3.19 Fitted E Horizon CEC vs. Measured and Fitted E Horizon CEC vs. Residuals

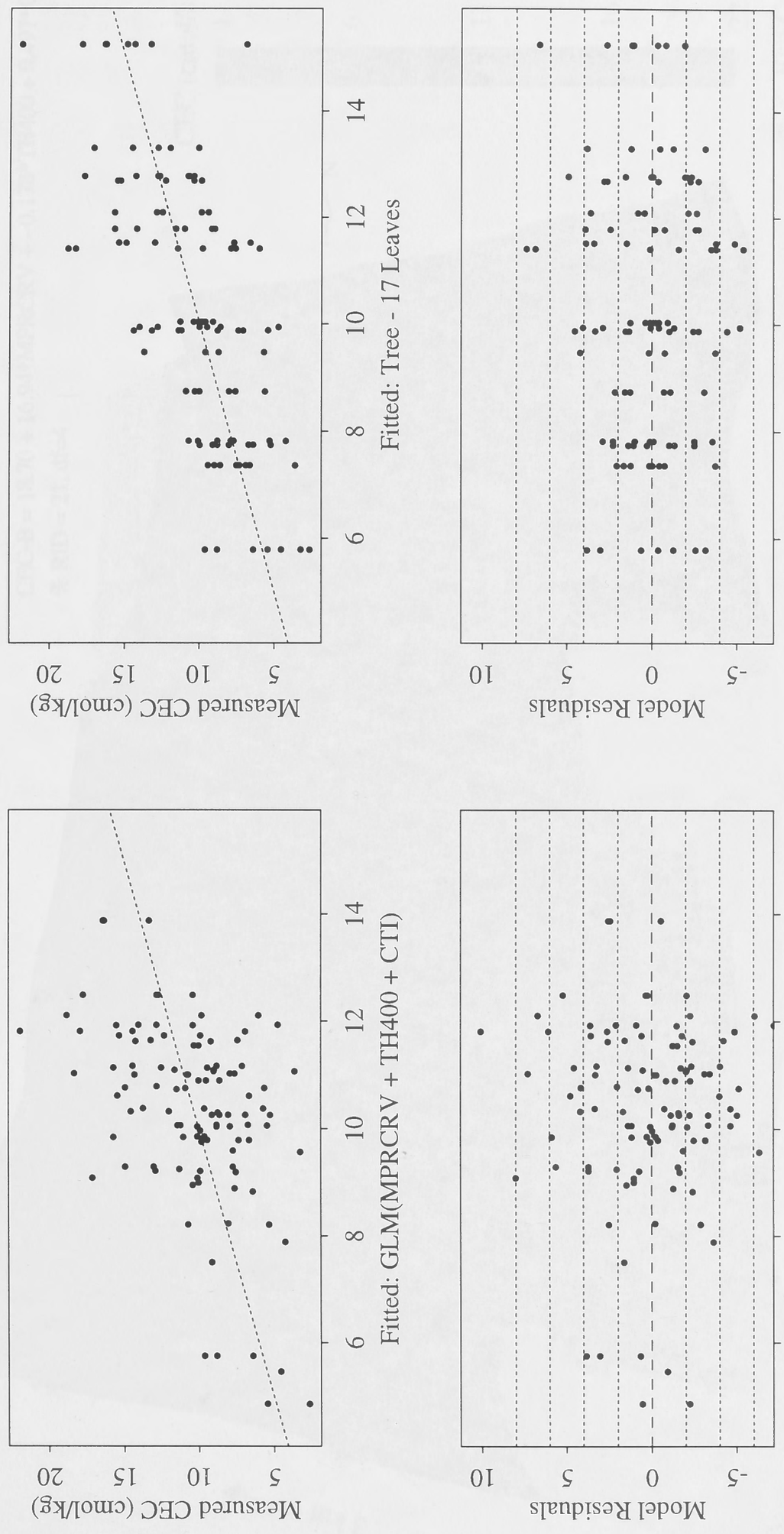


Figure 3.20 Fitted B Horizon CEC vs. Measured and Fitted B Horizon CEC vs. Residuals

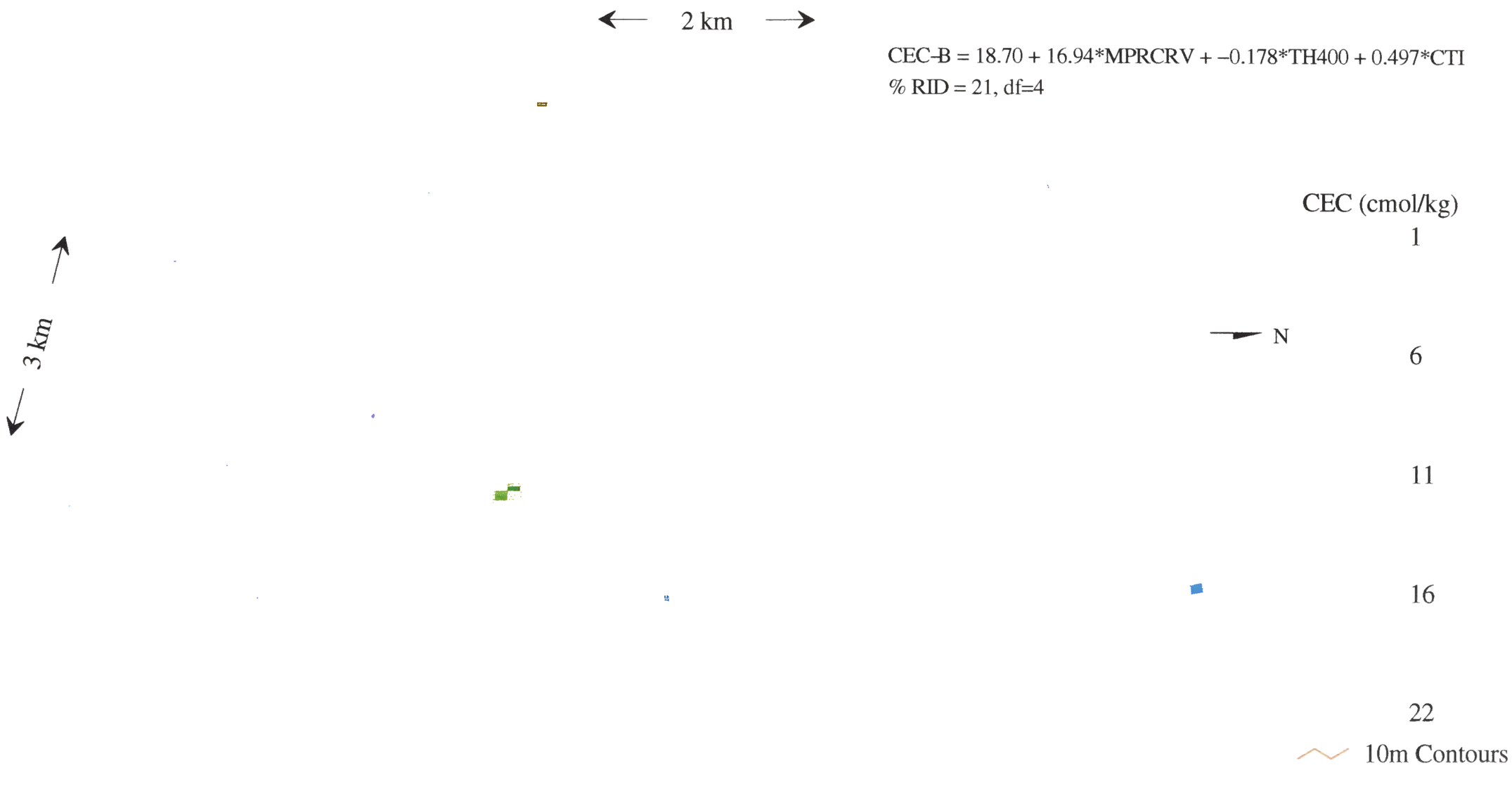


Figure 3.21 Drape of Predicted B Horizon Cation Exchange Capacity

sample depth, elevation, total annual precipitation, TC400, upslope mean profile curvature, K400, aspect and flow accumulation. The selection of sample depth as the most important explanatory variable for both the GLM and Tree B horizon CEC models may be suggesting two separate B horizon populations according to depth or a pedologic process such as clay illuviation depositing clays deeper in the profile or *in situ* clay formation. Inspection of Figures 3.15b and 3.16 does not show separate CEC clusters according to depth. Examination of the Ladysmith profile descriptions does indicate higher clay contents in the deeper B horizons suggesting a possible explanation.

Because sample depth was not available in a continuous manner over the study area, the step.gam procedure and Tree models were re-run without sample depth as a potential explanatory variable. The resulting Tree consumed 13 degrees of freedom with a %RID of 64 and produced 17 leaves. The step.gam procedure selected a three term GLM with a %RID of 21 using upslope mean profile curvature, TH400 and CTI. Figure 3.20 illustrates the fitted versus measured and fitted versus residuals for these two B horizon models. The normal error model appears appropriate, as was initially indicated by the B horizon CEC boxplot.

Spatial Prediction and Display

B horizons are often considered an important storehouse of nutrients important for plant growth. Therefore a spatial implementation of the B horizon CEC GLM model was developed (see Figure 3.21). The low level of certainty (%RID = 21) is reflected by the patchy or noisy nature of the predicted surface. Upslope mean profile curvature (MPRCRV) is computed using the d8 flow routing technique. This is the likely cause of the linear patterns of connected cells proceeding down slope on the predicted surface.

3.3.4 Exchangeable Sodium Percentage

Exchangeable sodium percentage (ESP) is a measure of the cation adsorption sites occupied by sodium per unit weight of soil. Soils with cation exchange sites

dominated by sodium disperse when wetted and structurally collapse causing low permeability and waterlogging leading to reduced biological activity, trafficability and erosion problems. In Australia, sodic horizons are broadly defined by Isbell (1995) as having an ESP of six or more.

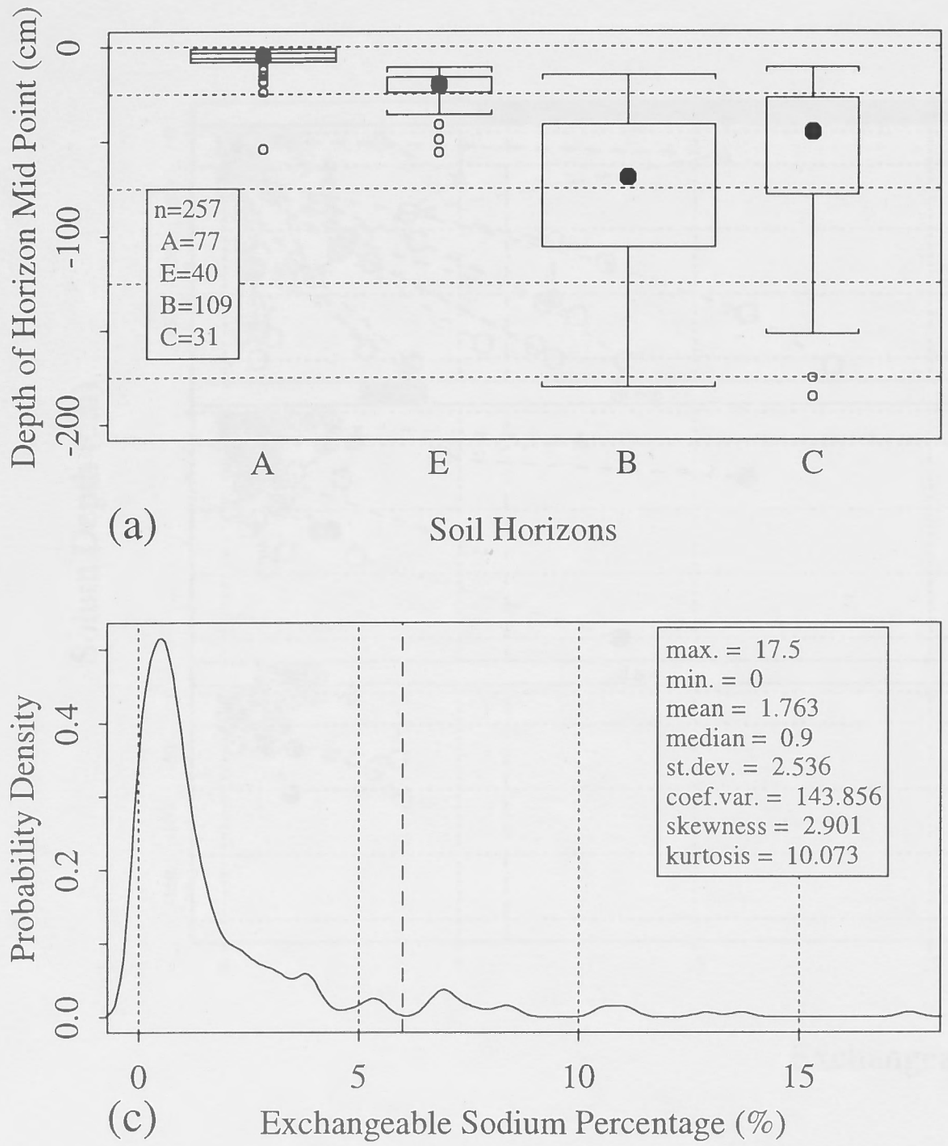
Exploratory Plots

Figure 3.22 shows the univariate and bivariate EDA plots for ESP. A vertical long-dashed line is plotted at ESP equal to six. The sample distribution is strongly peaked and positively skewed. Figure 3.22d indicates that ESP is generally low throughout the soil horizons with a scattering of sodic B horizons and two sodic E horizons. The horizon distributions are positively skewed. Figures 3.23 and 3.24 illustrate coplots of the ESP profiles conditioned by CTI, and slope and specific catchment area, respectively. These indicate that the two sodic E horizons are in profiles that contain other sodic horizons. Figure 3.24 shows a stronger clustering of sodic horizons towards the upper left of the coplot indicating a tendency towards low slope and large catchment area positions (i.e. bottom of the hillslope continuum). The coplots do not suggest a definable landscape threshold where sodic horizons begin to occur.

Given that sodium is soluble and very mobile (Hudson, 1995), this suggests that solute transport of sodium to the low parts of the landscape is an important process in local areas affected by sodicity. A plotting of those sample locations with sodic horizons showed a geographic clustering on the edge of the Ordovician metasediment geological unit in close proximity to the Kyeamba Creek valley floor (runs through the centre of the Ladysmith study area - Fig. 3.1).

Stepwise Attribute Selection and Model Development

Since ESP's greater than six were limited mostly to B horizons and the overall variability was very low, only B horizon ESP models were developed. Step.gam selected an eight linear term GLM consuming nine degrees of freedom giving a %RID of 56. The variables used in order of deviance reduction were sample depth, TH400,



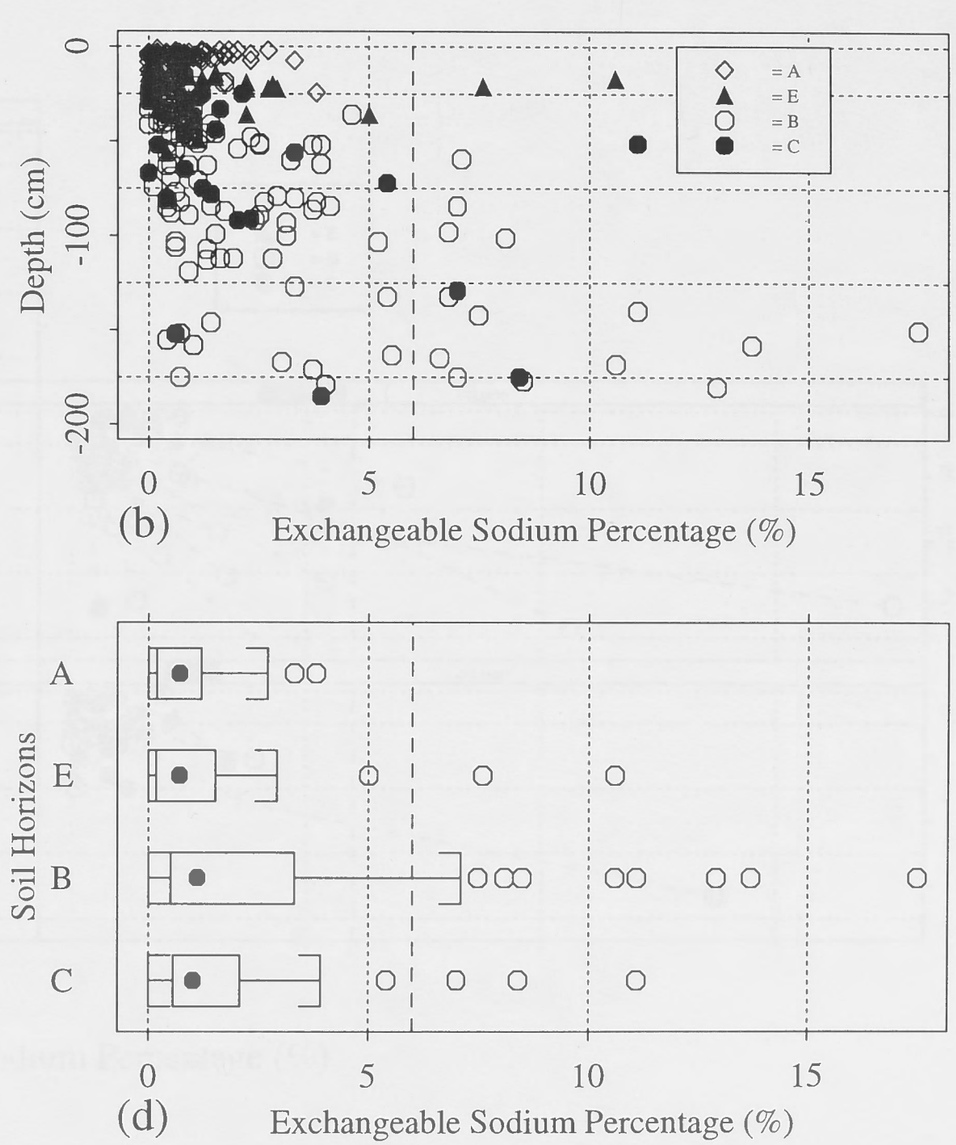


Figure 3.22 Exchangeable Sodium Percentage Univariate and Bivariate EDA (Ladysmith)

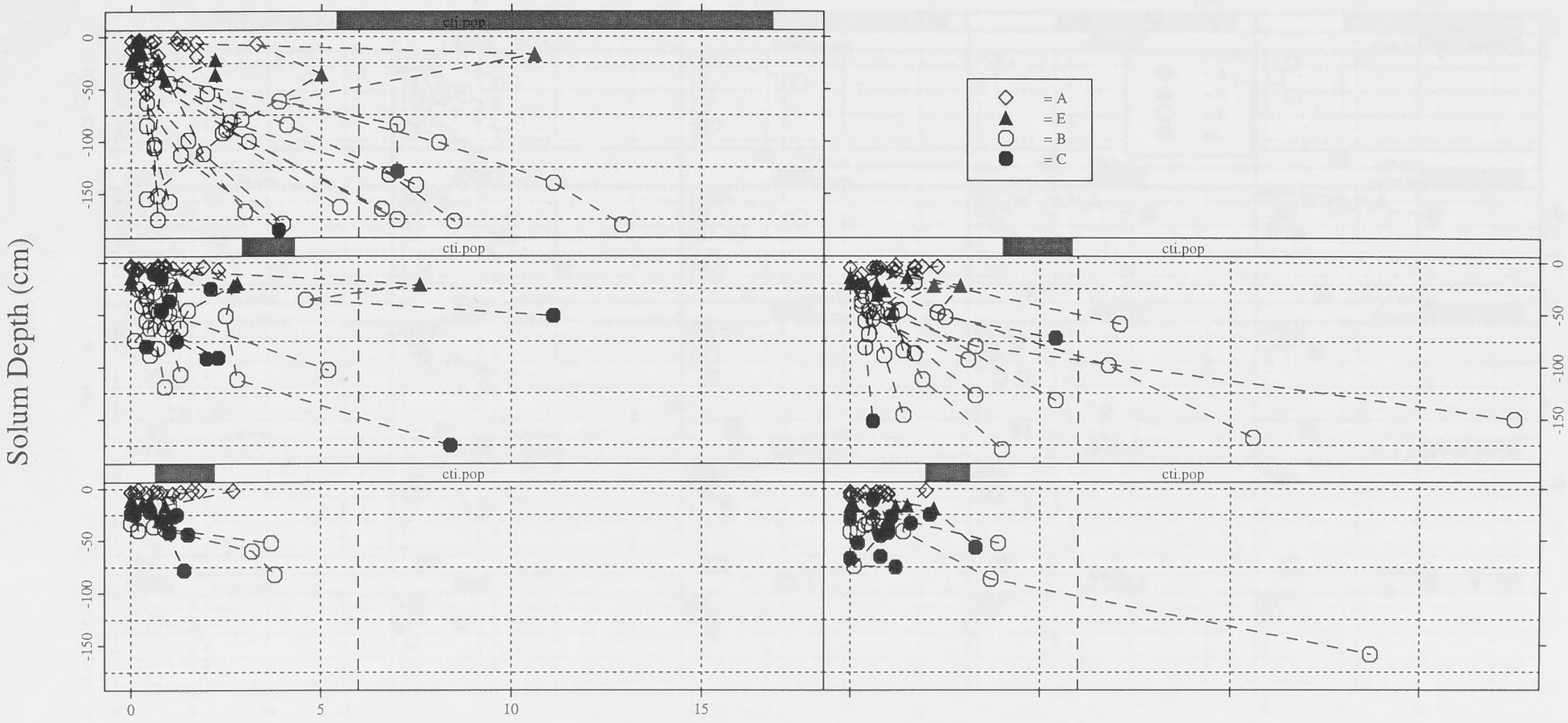


Figure 3.23 Exchangeable Sodium Percentage Coplot Conditioned by Compound Topographic Index (Ladysmith)

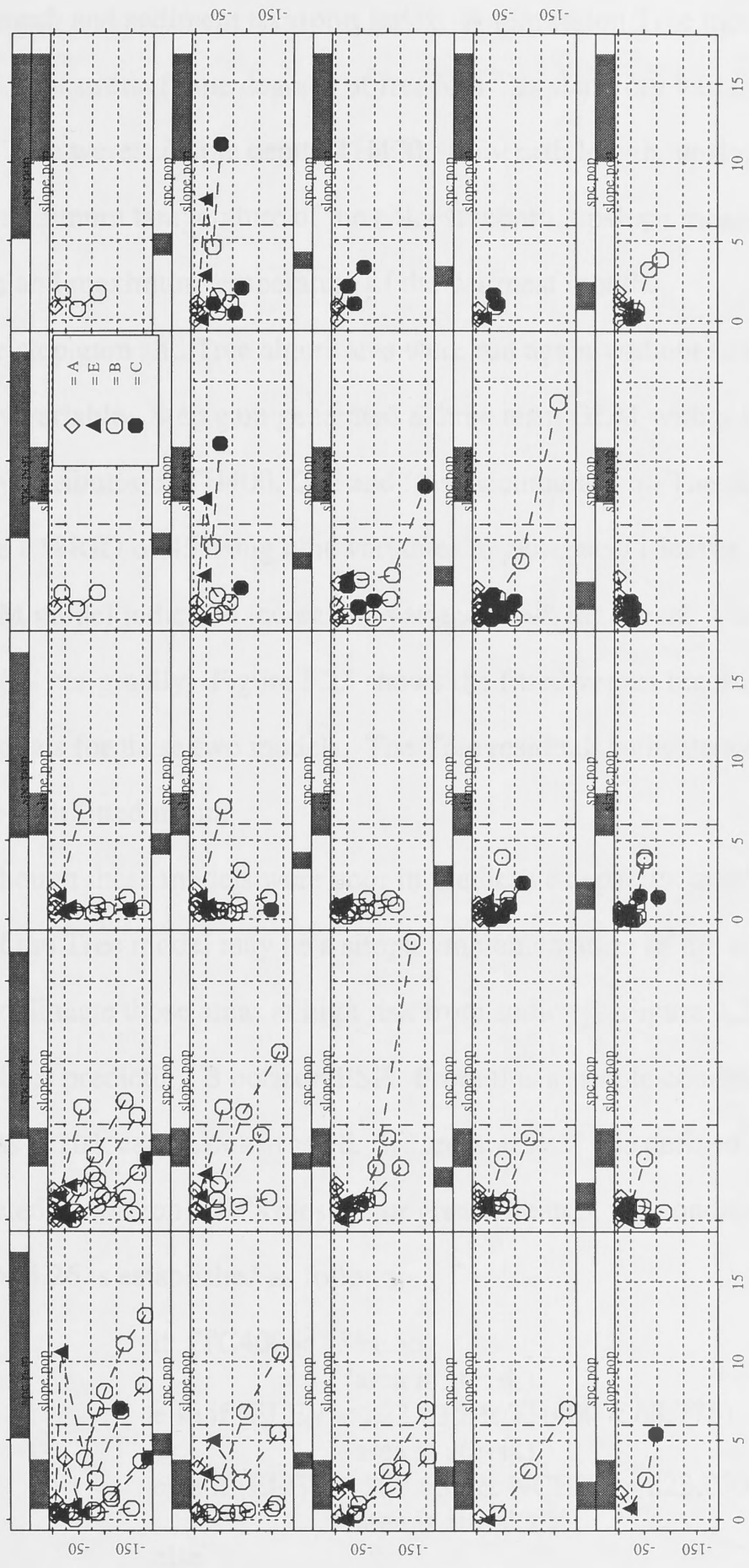


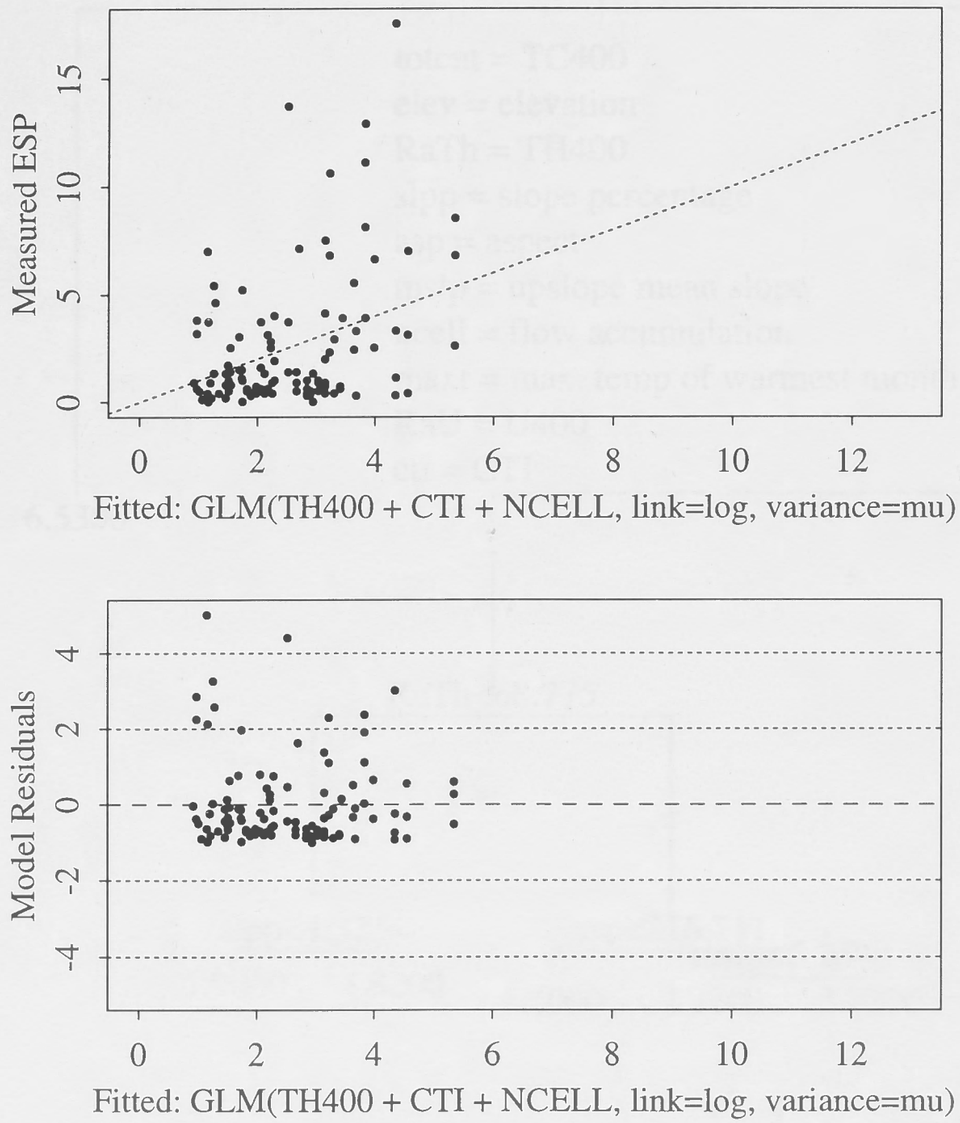
Figure 3.24 Exchangeable Sodium Percentage Coplot Conditioned by Slope and Specific Catchment Area

profile curvature, U400, flow accumulation, elevation, maximum temperature of the warmest month and sediment transport index. A regression Tree model provided a %RID of 81 consuming nine degrees of freedom. Explanatory variables used in the regression Tree were: sample depth, TH400, flow path length, upslope mean profile curvature, minimum temperature of the coldest month, upslope mean slope, flow accumulation and maximum temperature of the warmest month.

The step.gam and Tree algorithms were run again without sample depth as an explanatory variable. Step.gam generated a three term GLM with a 11 %RID using explanatory attributes of TH400, CTI and flow accumulation. The regression Tree model gave a %RID of 45 using nine variables to generate 11 leaves. The residuals for the GLM model indicated increasing variance with the mean. Use of the log link improved this marginally. Figure 3.25 shows the fitted versus measured and fitted versus residuals for these two models. The Tree residuals indicate a strong increase in variance with fitted mean.

Although these models were poor in predictive capacity, another potential application of the Tree model may be a simple implementation of the conditional rules defined to delineate those areas at high risk from sodicity. Figure 3.26 shows the Tree model for predicting B horizon ESP. From this a simple conditional rule to delineate areas at risk can be constructed. If "areas at risk" are defined as those areas with predicted B horizon ESP's close to or greater than 5, the conditional rule set from Figure 3.26 is established as follows:

```
if  (TC400 < 1551.56)
                (area is at risk)
else if (ELEV < 223.635 & TH400 < 68.775)
                (area is at risk)
else if (ELEV > 223.635 & NCELL > 223.033)
                (area is at risk)
else
                (area not at risk)
endif
```



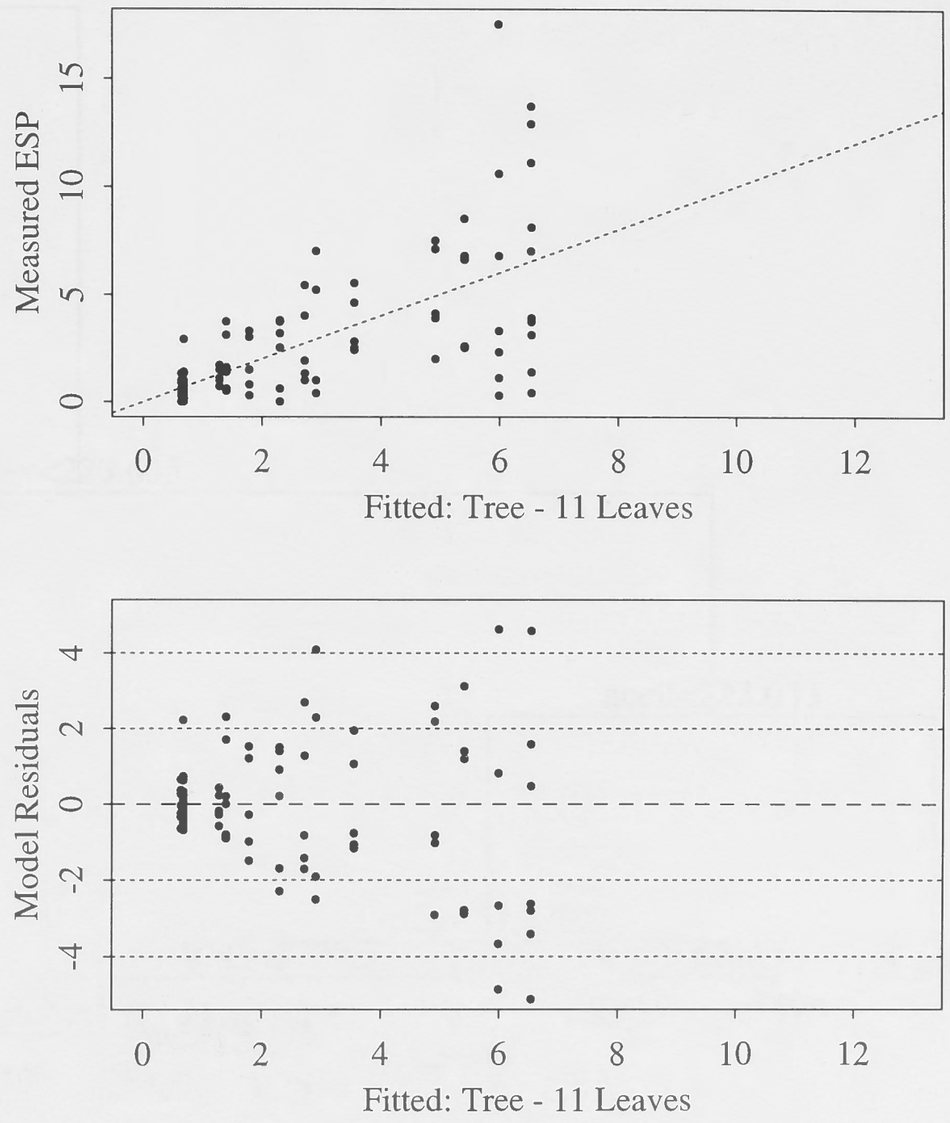


Figure 3.25 Fitted B Horizon ESP vs. Measured and Fitted B Horizon ESP vs. Residuals

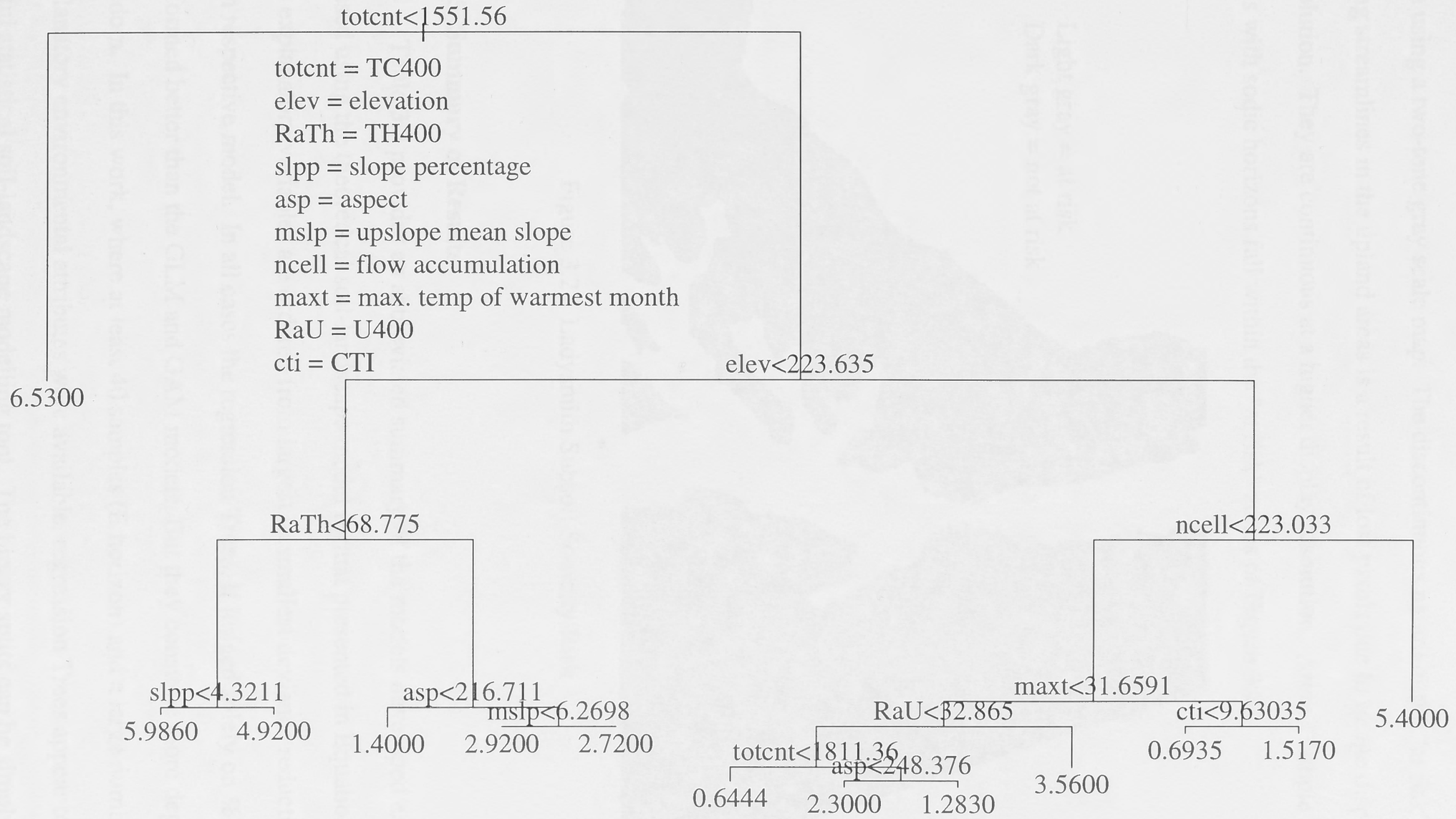


Figure 3.26 B Horizon ESP Regression Tree Model

Figure 3.27 displays a spatial implementation of this rule set for the Ladysmith study area using a two-tone gray scale map. The discontinuous nature of the "at risk" areas along streamlines in the upland areas is a result of lost pixels due to image display resolution. They are continuous at a higher display resolution. All the sample locations with sodic horizons fall within the "at risk" zones of Figure 3.27.

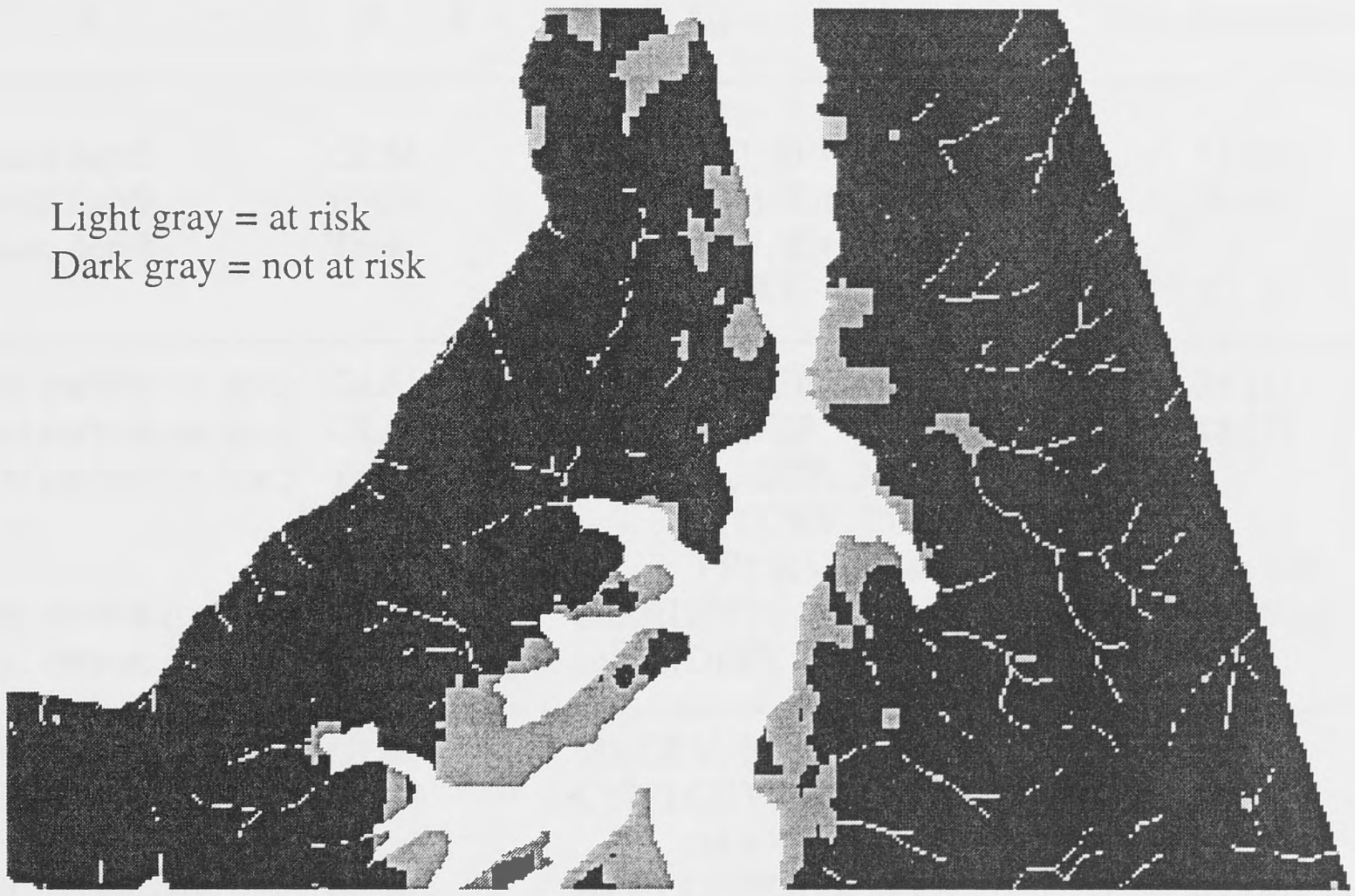


Figure 3.27 Ladysmith Subsoil Sodicity Risk

3.3.5 Summary of Results

Table 3.2 provides an abbreviated summary of the models developed expressed using the theoretical soil-landscape model format presented in Equation 2.1. The explanatory variables are ordered from largest to smallest deviance reductions for each respective model. In all cases the regression Trees, if judged solely on %RID, performed better than the GLM and GAM models. But they consume more degrees of freedom. In this work, where at least 40 samples (E horizon) and a large number of explanatory environmental attributes were available, regression Trees appear to be a useful statistical soil-landscape modelling tool. The binary splits can be simply

Table 3.2 Soil-landscape Models for Ladysmith Study Area

i_n = Ladysmith Ordovician Metasediment Environmental Domain

response variable **s**	~	model type f	(	explanatory environmental variables $x_1, \ldots x_n$	)	%RID (d.f. consumed)
solum depth		GLM		CTI, K400, SPI		75 (4)
solum depth		GAM		spline(CTI), K400		78 (6)
solum depth		Tree		CTI, ELEV, K400, MPRCRV, TCRV		90 (5)
total carbon (A hor.)		GLM		MTCRV, DEP		38 (3)
total carbon (A hor.)		$GLM_{(log\ link)}$		MTCRV, DEP		48 (3)
total carbon (A hor.)		Tree		CTI, FPL, DEP, ASP, MTCRV, PDQ, TCRV, PRCRV, TH400, SLPP		79 (10)
total carbon (profile)		$GAM_{(log\ link)}$		loess(DEP)		84 (6.8)
total carbon (profile)		$GAM_{(log\ link)}$		spline(DEP)		83 (6)
CEC (A horizon)		GLM		MTCRV, NCELL, U400		33 (4)
CEC (A horizon)		Tree		CTI, TCRV, MPRCRV, TH100, MAXT, U400, MSLP		74 (7)
CEC (E horizon)		GLM		NCELL, TH400, SPI		41 (4)
CEC (E horizon)		Tree		PDQ, TC400, DEP, ELEV, U400		45 (6)
CEC (B horizon)		GLM		MPRCRV, TH400, CTI		21 (4)
CEC (B horizon)		Tree		ELEV, PRCRV, MTCRV, STI, MSLP, MPLCRV, ASP, K400, CTI, MPRCRV, U400, PLCRV		45 (12)
ESP (B horizon)		$GLM_{(log\ link)}$		TH400, CTI, NCELL		11 (4)
ESP (B horizon)		Tree		TC400, ELEV, TH400, SLPP, ASP, MSLP, NCELL, MAXT, U400		45 (9)

converted to decision rules for spatial implementation and field evaluation. However, depending on the number of leaves and distribution in response attribute space, Tree models may produce a non-continous or stepped prediction surface on the landscape. In instances where discontinuous soil patterns exist this may be appropriate,

but in many cases it isn't. The GLM and GAM models do not suffer from this limitation. The capability to change error models depending on residual behavior appears to be another advantage of the GLM and GAM methods. This work indicates that the log link function improved residual patterns for total carbon and ESP. GAM's emerge as most appropriate when smooth, non-linear relationships exist between response and explanatory variables.

The exploratory plots were central to indicating when sample depth or horizon factors are significant stratifiers of variation and when transformations may be appropriate. The coplots indicated no sharp breaks or landscape thresholds according to CTI, slope and specific catchment area conditioners suggesting that the Ladysmith soil-landscapes are a continuum. The fitted versus measured and fitted versus residual plots provided a simple display for evaluation and comparison of models complementing the %RID values for each model. Likewise, colour visualizations also conveyed an understanding of differences between models and of model certainty.

A broad range of explanatory variables were useful for soil attribute prediction. Lower meso-scale terrain attributes were the most useful and the gamma radiometric variables provided a useful complement in several models (solum depth, CEC, ESP). The %RID's ranged from 11 to 90, showing that some attributes are more predictable than others using this sample dataset. The use of a flexible modelling approach that uses several tools to search for useful relationships was important because of the difference in variance characteristics for different variables. Each individual model may be analyzed in greater detail than demonstrated here with additional residual diagnostic plots and analysis of deviance tables to check interactions between model terms to further improve the model and interpretations. The intention was to demonstrate the mechanics of a statistical soil-landscape modelling approach using contemporary tools to develop explicit and quantitative models of individual soil attributes or a combination of models for varied purposes.

3.4 CONCLUSIONS

This chapter demonstrates that explicit and quantitative environmental correlations can be derived to spatially predict individual soil attributes using statistical models with stated levels of uncertainty and model complexity. Each step has been explicitly stated and quantitative models using various environmental correlations have been defined using a variety of exploratory data analysis and statistical modelling tools. This places the entire approach in a framework that can be tested, improved and adapted for various purposes.

A new sampling approach using a CTI provisional model and an integration of traditional and spatial statistical theory to allocate samples in geographic space was developed and applied. This enabled the visualization of soil attribute variation as it changes through the landscape allowing simple assessment of soil-landscape patterns and potential thresholds for soil attribute variation. Evidence from the models developed here suggests that the Ladysmith soil-landscapes are a continuum. This approach employs a very different philosophy compared to traditional transect sampling approaches. A hillslope transect provides useful information about the specific hillslope under study, but does not convey much about the hillslope patterns over a broad spatial area. The approach used here collects data over a gradient of hillslope positions from a broad spatial area to represent the concept of a spatially-averaged hillslope. This statistically-based approach provides a more appropriate representation of the variation in an area. Additional research is required to develop theories on the number of samples required in a spatial area. This will depend on many variables (e.g. soil variables to be measured, physiographic complexity of area, resources available).

Bias is entered into the sampling process by elimination of some areas from selection and the general practice of selecting the centre of a quantile patch. This violates the desire for random selection in that all grid nodes do not have an equal chance

of being selected. However, initial rough field positioning followed by more accurate positioning after sampling entered an approximate 40m positioning randomness based on intended location and post-sampling location analysis. This enters an element of randomness but does not negate the intended spreading of samples through CTI space because the quantile breaks were used only as a guide to sample in continuous CTI attribute space. Plots of the spread of samples in CTI and several other terrain attribute spaces after sampling showed a good distribution of samples in all attribute spaces.

The use of several exploratory tools and a stepwise explanatory attribute selection process enabled a dynamic and complementary search for useful environmental correlations. In instances where patterns were not apparent, more reliance was placed on the step.gam method to select useful explanatory variables. Although some simple landscape process interpretations were suggested, it may be better to make decisions on explanatory variables used based on hypothesized landscape process interpretations. However, because variables can only be measured at limited space-time scales, decisions based solely on process interpretations may limit the capacity to predict. An approach that seeks to initially utilize both may be best.

The preservation of both sample depth and horizon classification factors proved useful for microscale stratification and prediction. This enables greater flexibility in the development of models of specific attributes or for integrated models of variation over the entire profile. Model %RID's ranged from 11 (B horizon ESP GLM) to 90 (solum depth Tree) indicating a broad range of predictability. Webster (1977) concluded that the variation accounted for by a typical general purpose soil survey would range from about half the total variance for soil physical attributes to less than one tenth for some chemical attributes. This provides an informal measure for comparison that suggests the models developed here are good, considering the general noisy nature of soil data. Spatial predictions that use lower meso-scale

explanatory terrain attributes (20m) provide maps at a high spatial resolution that should prove useful for applications at the land management level.

The lack of a good model or strong predictive relationships may imply that relationships do not exist, but it may also mean that the empirical evidence gathered does not support such an assertion. Lack of relationships may be due to a mismatch in scales of measured soil and environmental variables or other underlying causes that are not captured by the suite of environmental variables used. The use of this explicit and quantitative approach present no bounds to continued analysis and redevelopment of models using additonal data or other modelling aproaches. The framework is constructed with a view that resource inventory and generation of derived products should not be a static process, but established in a manner that encourages continued analysis, testing and development using GIS and statistical modelling tools with additional data, knowledge and environmental variables.

3.5 REFERENCES CITED

Bevan, K.J., and M.J. Kirkby. 1979. A physically-based variable contributing area model of basin hydrology. Hydro. Sci. Bull. 24:43-69.

Bierwirth, P., P.E. Gessler, and D.J. McKane. 1996. Empirical investigation of airborne gamma-ray images as an indicator of soil properties - Wagga Wagga, NSW. Proceedings of the 8th Australasian Remote Sensing Conference. Canberra.

Butler, B.E. 1956. Parna - an aeolian clay. Australian J. of Sci. 18:145-151.

Chen, X.Y., and D.J. McKane. in prep. Soil-landscapes of the Wagga Wagga 1:100 000 sheet and the Kyeamba Valley. NSW Conservation & Land Management. Sydney, Australia.

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, New Jersey.

Gallant, J.C. 1996. TAPES terrain analysis programs. World Wide Web. http://cres.anu.edu.au/software/tapes.html.

Geeves, G.W., H.P. Cresswell, B.W. Murphy, P.E. Gessler, C.J. Chartres, I.P. Little, and G.M. Bowman. 1995. The physical, chemical and morphological properties of soils in the wheat-belt of southern N.S.W and northern Victoria. CSIRO Division of Soils Report. Adelaide, Australia.

Gessler, P.E., and Ashton, L.J. in prep. Wagga Wagga geographical information system database: development, structure and user access. CSIRO Divisional Working Report No.X CSIRO Division of Soils. Canberra.

Heuvelink, G.B.M. 1993. Error propagation in quantitative spatial modelling. Ph.D. Dissertation. University of Utrecht, Utrecht, The Netherlands.

Hudson, B.D. 1995. Reassessment of Polynov's ion mobility series. Soil Sci. Soc. Am. J. 59:1101-1103.

Hunter, G.J., and M.F. Goodchild. 1995. Dealing with error in spatial databases: a simple case study. Photogrammetric Engineering and Remote Sensing. 61(5):529-537.

Hutchinson, M.F. 1989. A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. J. of Hydrol. 106:211-232.

Hutka, J. 1994. Sedigraph 5100 particle size system: a brief desription and its use in soil particle size analysis work. CSIRO Divison of Soils Technical Report 9/1994. Canberra, Australia.

Hutka, J. and L.J. Ashton. 1995. Sedigraph particle size analysis of soil samples from the Griggward study area of the Wagga Wagga project. CSIRO Division of Soils Technical Report 3/1995. Canberra, Australia.

Isbell, R.F. 1995. The Australian soil classification system. Australian Soil and Land Survey Handbook, no. 4. Canberra, Australia.

McCullagh, P., and J.A. Nelder. 1989. Generalized linear models. 2nd Ed. Chapman and Hall, London.

McDonald, R.C., Isbell, R.F., Speight, J.G., Walker, J., and Hopkins, M.S., 1990. Australian soil and land survey field handbook. Second Edition. Inkata Press, Sydney.

McMahon, J.P., M.F. Hutchinson, H.A. Nix, and K.D. Ord. 1995. ANUCLIM: users guide. Centre for Resource & Environmental Studies, Australian National University. Canberra, Australia.

Merry, R.H. and L.R. Spouncer. 1988. The measurement of carbon in soils using a microprocessor-controlled resistance furnace. Commun. Soil Sci. Plant Anal. 19(6):707-720.

Moore, I.D. 1992. Terrain analysis programs for the environmental sciences (TAPES). Agric. Sys. Inf. Tech. 4(2):37-39.

Moore, I.D., A.R. Ladson, and R. Grayson 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydro. Proc. 5:3-30.

Moore, I.D., P.E. Gessler, G.A. Neilsen, and G.A. Peterson. 1993a. Soil attribute

prediction using terrain analysis. Soil Sci. Soc. Am. J. 57:443-452.

Moore, I.D., A. Lewis, and J.C. Gallant. 1993b, Terrain attributes: estimation methods and scale effects. p. 189-214. *In* A.J. Jakeman (ed.) Modelling Change in Environmental Systems. London, Wiley.

Milne, G. 1935. Some suggested units of classification and mapping particularly for East African soils. Soil Research 4:3.

Raymond, O.L., 1992. Geology of the Wagga Wagga sheet. Sheet 8327, Australian Geological Survey Organization, Canberra, Australia.

Searle, P.L. 1986. The measurement of soil cation exchange properties using the silver thiourea method: an evaluation using a range of New Zealand soils. Aust. J. Soil Res. 24:193-200.

Speight, J.G. 1968. Parametric description of landform. p239-250. *In* G.A. Stewart (ed.) Land Evaluation. Macmillan, Melbourne.

Speight, J.G. 1974. A parametric approach to landform regions. Special Publ. no. 7. Institue of British Geographers.

Statistical Sciences. 1993. S-PLUS Guide to statistical and mathematical analysis. Version 3.2. StatSci, a Division of MathSoft Inc., Seattle.

Trimble Navigation Limited. 1992. GPSurvey Software Users Guide. Trimble Navigation Limited. Sunnyvale, CA.

Webster, R. 1977. Quantitative and numerical methods in soil classification and survey. Clarendon Press, Oxford.

Chapter Four: Empirical Evaluation of DEM Resolution

4.1 INTRODUCTION

4.1.1 Broad Principles

The results of Chapter Three indicate that several quantitative terrain attributes can be used to spatially predict soil attributes. Recent studies show that grid based quantitative terrain attributes are scale dependent (Hutchinson and Dowling, 1991; Jenson, 1991; Panuska *et al.* 1991; Quinn *et al.* 1991; Moore *et al.* 1991; Zhang and Montgomery, 1994; Moore *et al.* 1994; Band and Moore, 1995; Bloschl and Sivapalan, 1995). Others suggest that certain scales may exist where environmental correlations, and hence predictive potential, are improved (Gerrard, 1990; Allen and Hoekstra, 1992; McSweeney *et al.* 1994; Band and Moore, 1995).

Recent studies evaluating the accuracy of terrain attribute computations conclude that much depends on the data source, compilation methods and computation techniques used to generate the DEM and derivatives (Lee *et al.*, 1992; Bolstad and Stowe, 1994; Brown and Bara, 1994; Fryer *et al.* 1994; Moore *et al.* 1994; McSweeney *et al.* 1994; Hammer *et al.* 1995; Gallant and Hutchinson, 1995). In a study focussing specifically on varied grid point spacing influences on the portrayal of the land surface for hydrologic simulations, Zhang and Montgomery (1994) conclude that a 10m grid spacing is a good compromise between the inaccuracy of larger grid sizes (30m and 90m) and the marginal improvements of smaller grids (2m and 4m) for two small study catchments. Moore *et al.* (1994) and Bloschl and Sivapalan (1995) suggest that more empirical studies are required that explicitly include definition of the physiographic context of the study (e.g. parent materials, local relief, dissection, landform patterns) to facilitate comparison and extrapolation.

The literature indicates that there are two broadly intertwined issues influencing quantitative terrain representation at varied scales. They are:

- influences relating to the source data, data structure and computational methods used; and
- influences relating to the landform variation and patterns in a particular physiographic domain (e.g. within a particular i_n environmental stratification).

Many studies do not provide explicit details or control over these influences to facilitate development of an understanding of environmental correlations and knowledge bases for application of terrain analysis methods at various scales in various settings.

In discussion of the potential use of joint probability distribution functions for scaling and extending spatial information for hydrological model parameterization, Band and Moore (1995) state: "A distinction needs to be made between simple data combination by GIS overlay and the extraction and synthesis of the spatial statistical associations within an area". Aside from the abundance of recent geostatistical work, the literature does not demonstrate methods for extracting and synthesizing spatial statistical associations within an area. Furthermore, rarely are spatially distributed field data used to evaluate environmental correlations and changes in predictive potential with measurement scales.

Discussion of spatial statistical associations in this context (Band and Moore, 1995) may be considered equivalent to environmental correlations as defined in Chapter One if the definition is expanded to include not only quantitative relationships between different environmental variables, but also quantitative relationships between the same environmental variable measured at different spatio-temporal scales.

For soil-landscape modelling, several basic questions must be addressed.

- How do terrain attributes change with differing grid point spacing?
- How does the potential for predicting soil attributes measured from soil cores change with different grid point spacings?

- Do certain scales exist where predictions are better or where certain terrain attributes are more useful?
- Does a more detailed topographic data source provide more predictive power?
- Does a scale exist where the resolution of the source data becomes less important?

This Chapter compares terrain attribute distributions measured at various meso-scale grid point spacings (5-80m) and evaluates empirical correlations between the varied scale terrain attributes and soil attributes measured from field core samples.

4.1.2 Hypotheses and Concepts

The three hypotheses tested in this chapter are:

- quantitative terrain attributes change systematically with scale;
- certain terrain attribute grid point resolutions exist where soil attribute prediction is better (as measured by %RID); and
- more detailed topographic data sources, closer to the scale of the soil attribute sample measurements, provide better predictions.

The significance of the first hypothesis is that systematic changes could be modelled and used to develop a scaling theory for use at different scales within a defined physiographic domain. The significance of the second and third hypotheses is that we should attempt to develop DEM's and derivatives at scales that balance data quantity and optimize predictive potential. These hypotheses are tested through the use of exploratory data analysis and statistical modelling techniques in a study area with a well-defined physiographic domain (i.e. fixed i_n) using a fixed set of computational methods.

4.2 MATERIAL & METHODS

4.2.1 Study Area - Griggward

Griggward (see figure 3.1), the first study area, was used for initial evaluation of topographic data sources. Preliminary terrrain analysis (Appendix Two; Gessler *et*

al. 1995) using DEM data derived from the 1:25 000 topographic map series indicated the study area incorporated a broad range of variation for the primary, secondary terrain attributes compared to previous work at other locations (Moore *et al.* 1993). This supported an intention to select an area capturing the landform variation typical of the broader Ordovician metasediment physiographic domain. A sub-area (17 km^2) was chosen for experimentation with DEM's at varied resolutions and from separate sources at 1:25 000 and 1:10 000 cartographic scales.

Figure 4.1 shows an orthophoto drape of the study area. This visualization conveys the range of landforms and drainage patterns characteristic of the Ordovician metasediment landscapes. Figure 4.1 complements the more regional hillshade for Griggward shown in Figure 3.1.

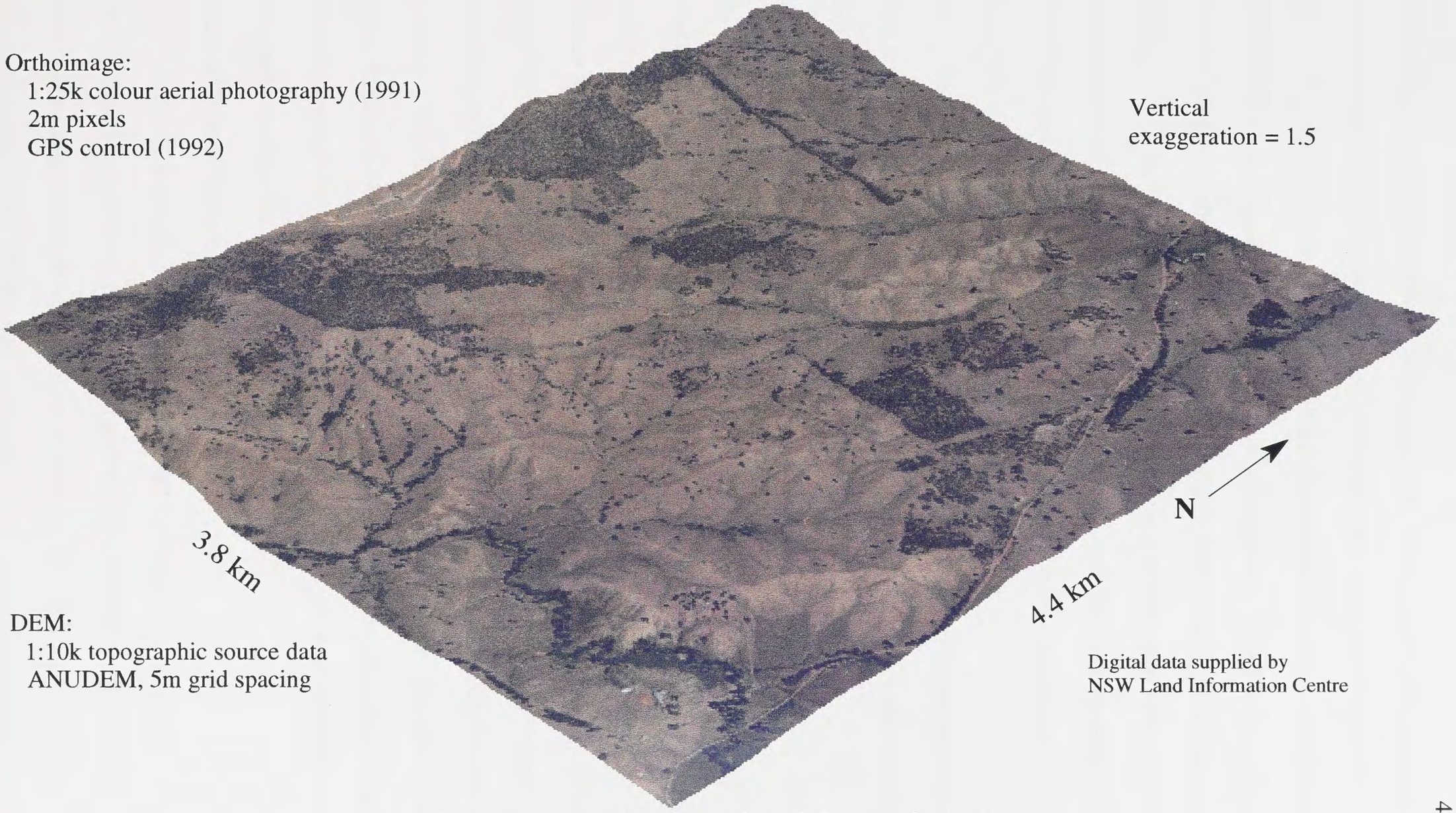
4.2.2 1:25 000 Source Topographic Data

Digital contours (10m), streamlines (x,y coordinate pairs) and spot heights (x,y,z coordinates) registered to the Australian Map Grid (AMG-UTM) were provided by the New South Wales Land Information Centre (NSW-LIC) in digital form. These data were generated photogrammetrically by a stereoplotter using 1:80 000 scale black and white aerial photography flown in 1986 with geopositioning horizontal and vertical control from the New South Wales Geodetic Survey. These topographic data are widely available in New South Wales as part of the 1:25 000 topographic map series.

4.2.3 1:10 000 Source Topographic Data

Digital contours (5m), streamlines (x,y,z coordinate pairs) and spot heights (x,y,z coordinates) registered to the Australian Map Grid (AMG-UTM) were provided by the New South Wales Land Information Centre in digital form. These data were generated photogrammetrically for this project by a stereoplotter using 1:25 000 colour aerial photography flown in February of 1991 (Mitchell, pers. comm.). Geopositioning horizontal and vertical control was developed by occupation of fence-post target control sites in conjunction with occupation of New South Wales

Orthoimage:
1:25k colour aerial photography (1991)
2m pixels
GPS control (1992)



DEM:
1:10k topographic source data
ANUDEM, 5m grid spacing

Digital data supplied by
NSW Land Information Centre

Figure 4.1 Orthophoto Drape Over Griggward Study Area

Geodetic Survey trigonometrical stations using two Trimble 4000 SSE Geodetic Basestation global positioning satellite (GPS) receivers (Gessler and Ashton, in prep.). The control points were post-processed for differential correction to provide accurate AMG coordinates for establishing the stereomodel. The photogrammetric compilation generated a large number of spot heights along ridge-tops, stream-lines and fence-lines to supplement the five metre contours providing a more detailed representation of the terrain. These are not generally available from standard 1:25 000 topographic maps. A digital orthophoto (2m pixels) was generated as a by-product of the photogrammetric work.

The control network design, field GPS collection and differential processing were done by the author. The calculation of trigonometrical station coordinates and photogrammetric work were performed by the NSW-LIC.

4.2.4 Soil Core GPS Positioning

The sampling strategy used for selection of soil core sample locations is described in section 3.2 and Appendix Two. After soil core collection, each location was occupied by a geodetic basestation (Trimble 4000 SSE) for five minutes in concert with geodetic basestation occupation of New South Wales Geodetic Survey trigonometrical stations. Differential post-processing was performed to provide accurate sample site coordinates. The GPS manufacturer indicates that this process should generate locations accurate to within a centimetre (Trimble Navigation Ltd., 1992). This process differs from the general process used in the other study areas in that two geodetic receivers were used instead of a geodetic and handheld receiver as previously discussed in Section 3.2.3.

4.2.5 DEM and Terrain Attribute Generation

ANUDEM (Hutchinson, 1988; 1989; 1995) was used for generation of all DEM's directly from the source topographic data. Four DEM's were generated at grid point spacings of ten, twenty, forty and eighty metres from the 1:25 000 scale topographic data. Five DEM's were generated at grid point spacings of five, ten,

twenty, forty and eighty metres from the 1:10 000 scale topographic data. These resolutions were chosen to:

- push the topographic source data to resolution limits (at small grid spacings);
- span meso-scales and hence data-set sizes feasible for implementation in a regional GIS;
- provide DEM's at equivalent resolutions to compare data sources;
- span meso-scales useful for provision of spatial predictions at the local hillslope scales; and
- span scales that may relate to meso-scale landscape processes.

All settings in ANUDEM followed the recommendations in the documentation.

TAPESG v5.0 (Moore, 1992; Gallant, 1996) was used for generation of all primary and secondary terrain attributes. Identical parameter settings were used for each run of TAPESG, with the exception of the bounding area which was altered slightly with each resolution to account for differing cell sizes. The multiple drainage direction algorithm was used with a maximum cross-grading threshold area of 30 000m^2. This defines the flow accumulation area where drainage direction switches from dispersive upland flow to D8 or channelized flow. The finite difference slope computation algorithm was used (Moore *et al.* 1993). The terrain attributes computed for this work were:

- elevation (DEM source data)
- slope gradient (%) (primary attribute);
- plan curvature (primary attribute);
- profile curvature (primary attribute);
- specific catchment area (secondary attribute);
- compound topographic index (secondary attribute);
- upslope mean slope (%) (secondary contextual attribute);
- upslope mean plan curvature (secondary contextual attribute);
- upslope mean profile curvature (secondary contextual attribute).

These attributes represent a range of DEM computational derivatives that quantify landform geometry and relate to various landscape processes as discussed by Speight

(1968;1974), Moore *et al.* (1991; 1993), McSweeney *et al.* (1994) and Gessler *et al.* (1995)(see Table 2.1).

Each terrain attribute was generated at each of the nine grid point resolutions for the same spatial area providing 81 grids ranging in population size from 670 441 values for the 5m resolution to 2 793 values for the 80m resolution. The grids were then cut out to correspond with the bounds of the Ordovician metasediment geology map unit. This eliminated some of the area around the southern edge of Figure 4.1 as shown by the hillshade in Figure 3.1.

A three component prefix/suffix naming convention is used to refer to the data source, grid point resolution and computed terrain attribute in the following sections. For example, c05g20.slopep, refers to a grid of percent slope (slopep) generated from the five metre contour (c05) source data (1:10 000 scale) at a grid resolution of twenty metres (g20). The c10g20 grids were used for soil-landscape modelling work in Chapters Three, Five and Appendix Two. Figure 4.2 illustrates DEM hillshades and the digital orthophoto for the 1:25 000 scale data. Figure 4.3 illustrates DEM hillshades and the digital orthophoto for the 1:10 000 scale data. These convey a visual impression of how resolution changes the definition of hillslopes over the landscape at the varied grid spacing resolutions. The 5m spacing DEM captures fine details such as streamline gulleys whereas the 80m grid spacing DEM only captures gross landform morphology.

4.2.6 Database Development

Two databases were developed. The first contained the values of each terrain attribute at each resolution for every cell over the study area. This database was used for comparison of terrain attribute distributions (pdf's). A second database was established to hold the soil layer data described from each soil core and each of the corresponding sample location terrain attributes at each resolution. The sample location coordinates were used to extract sample location terrain attributes from all the varied scale terrain attribute grids. This assumes that the grid point values are

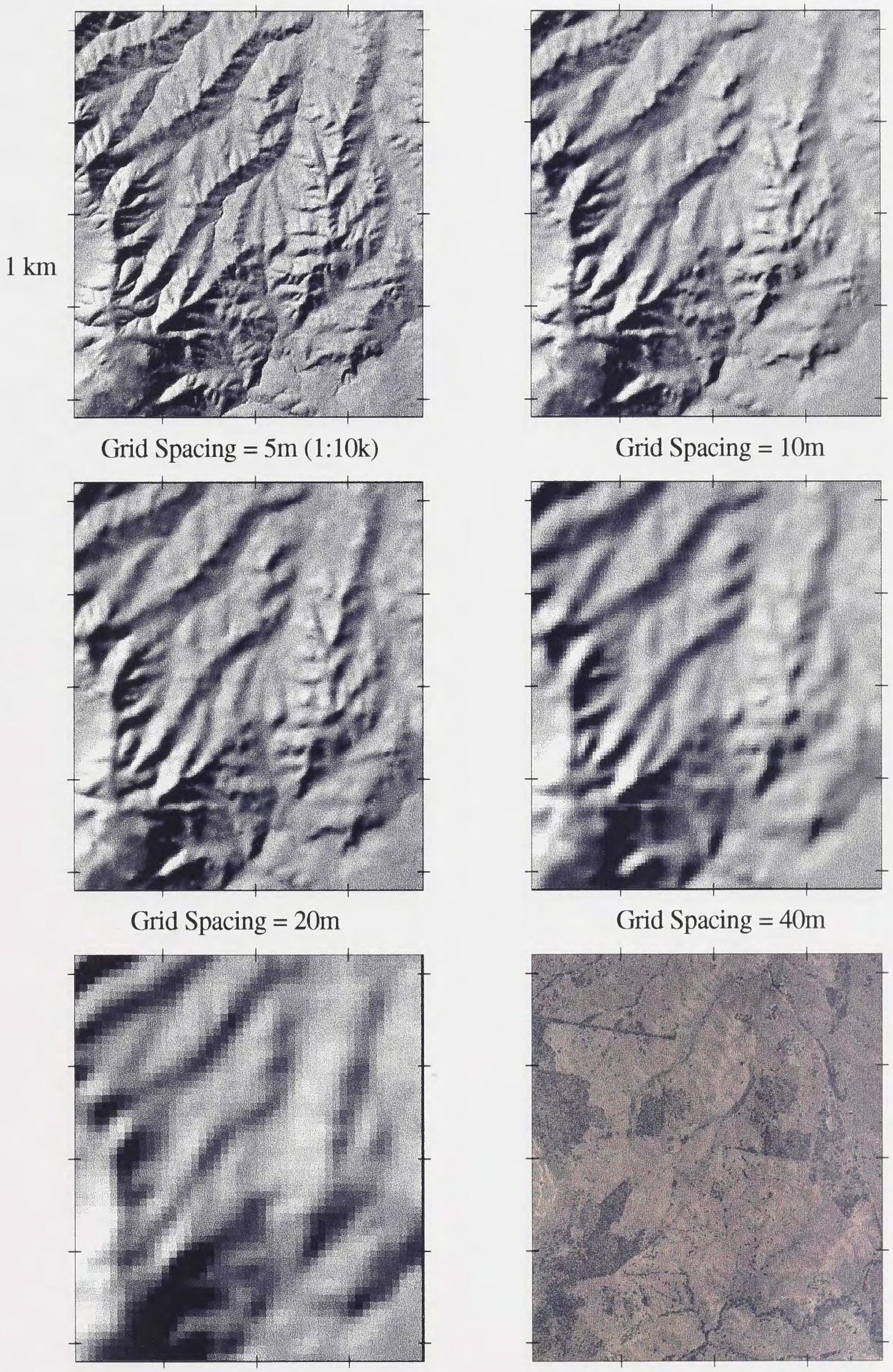


Figure 4.2 1:25 000 Source DEM Hillshades

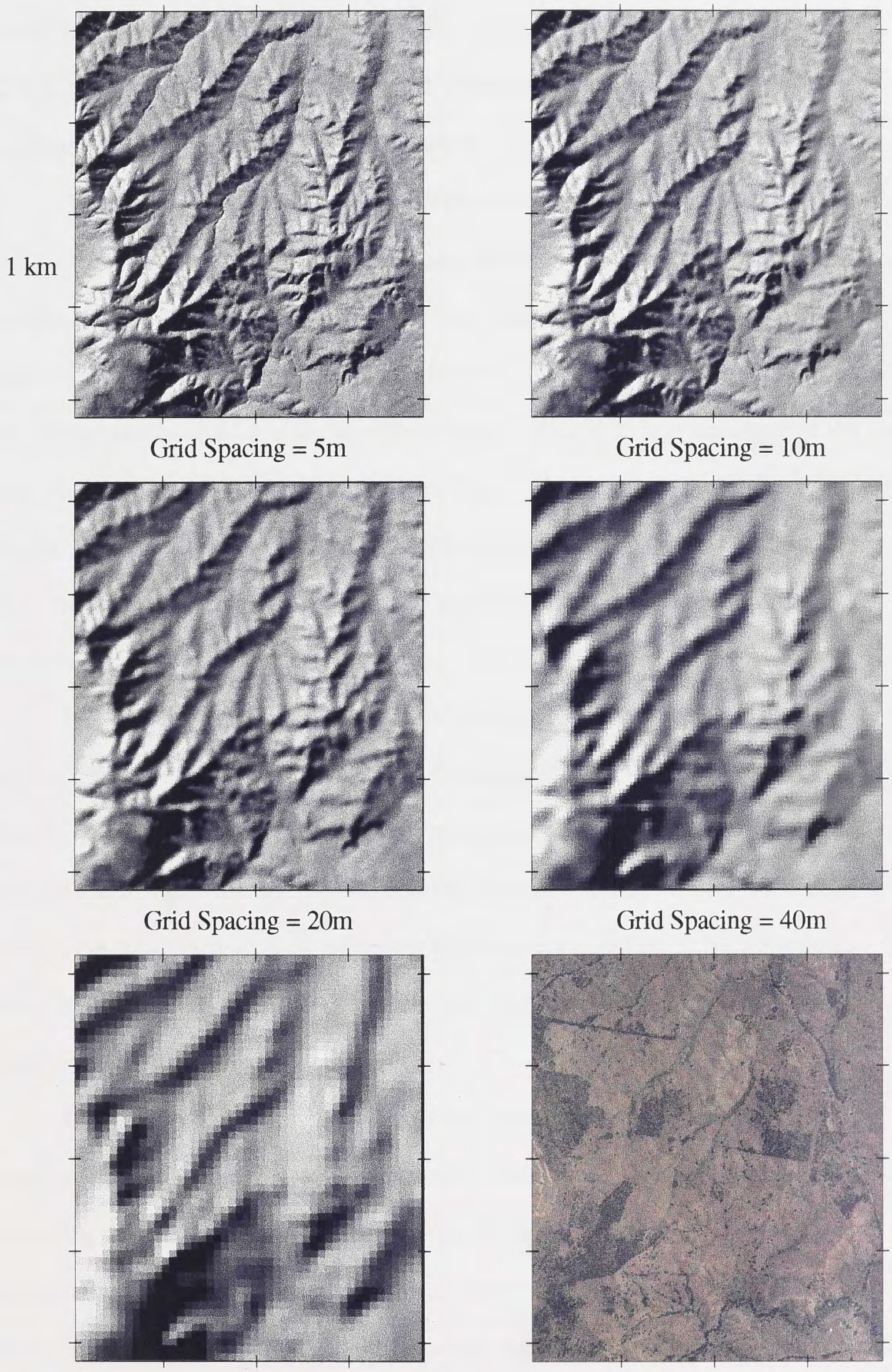


Figure 4.3 1:10 000 Source DEM Hillshades

representative of a square cell with dimensions equivalent to the respective grid spacing sizes (i.e. 5m-80m) around each grid point. The second database was used for soil layer prediction as discussed below.

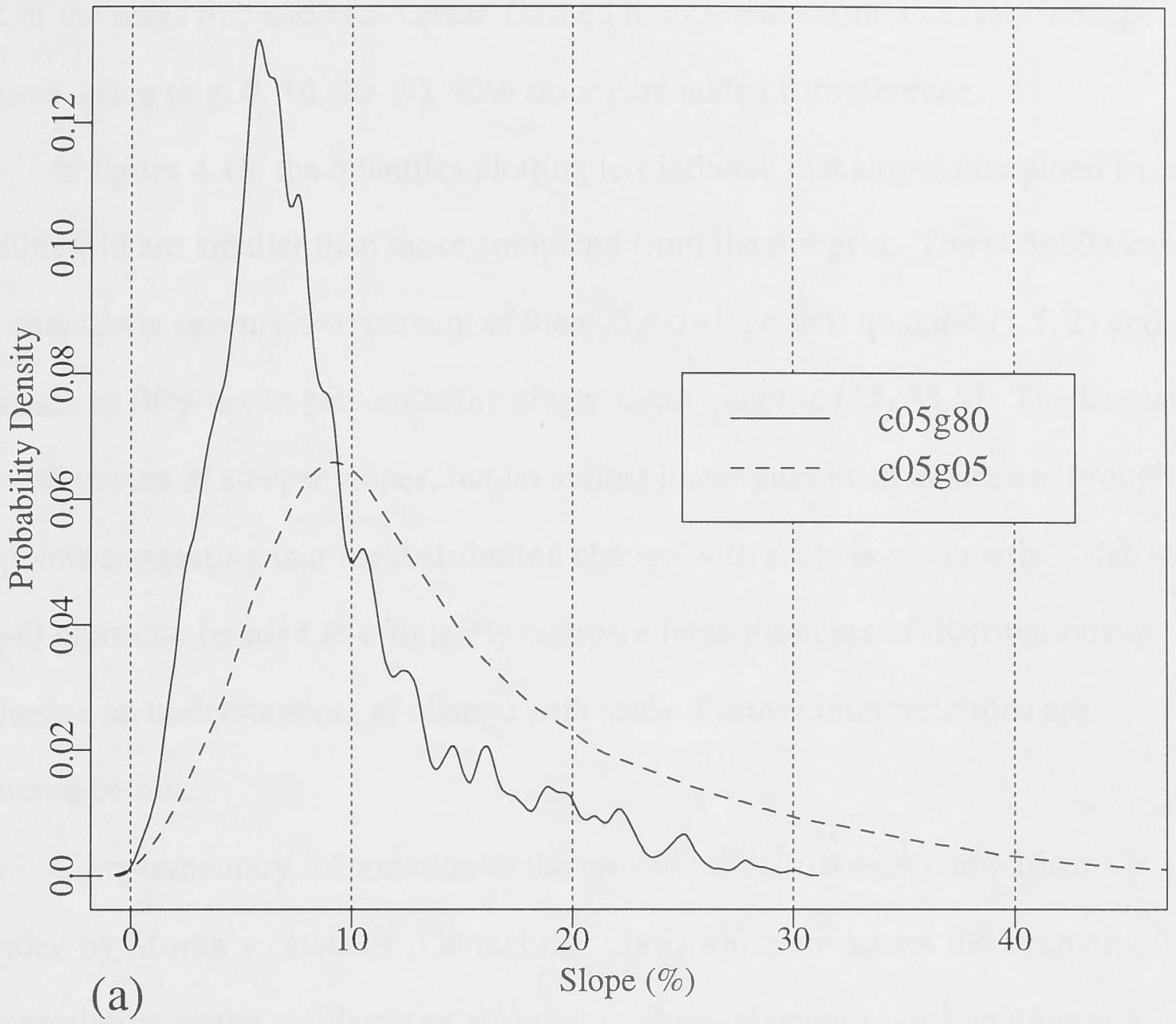
4.2.7 Comparison of Distributions and Predictive Utility

Empirical Comparison of DEM Resolution

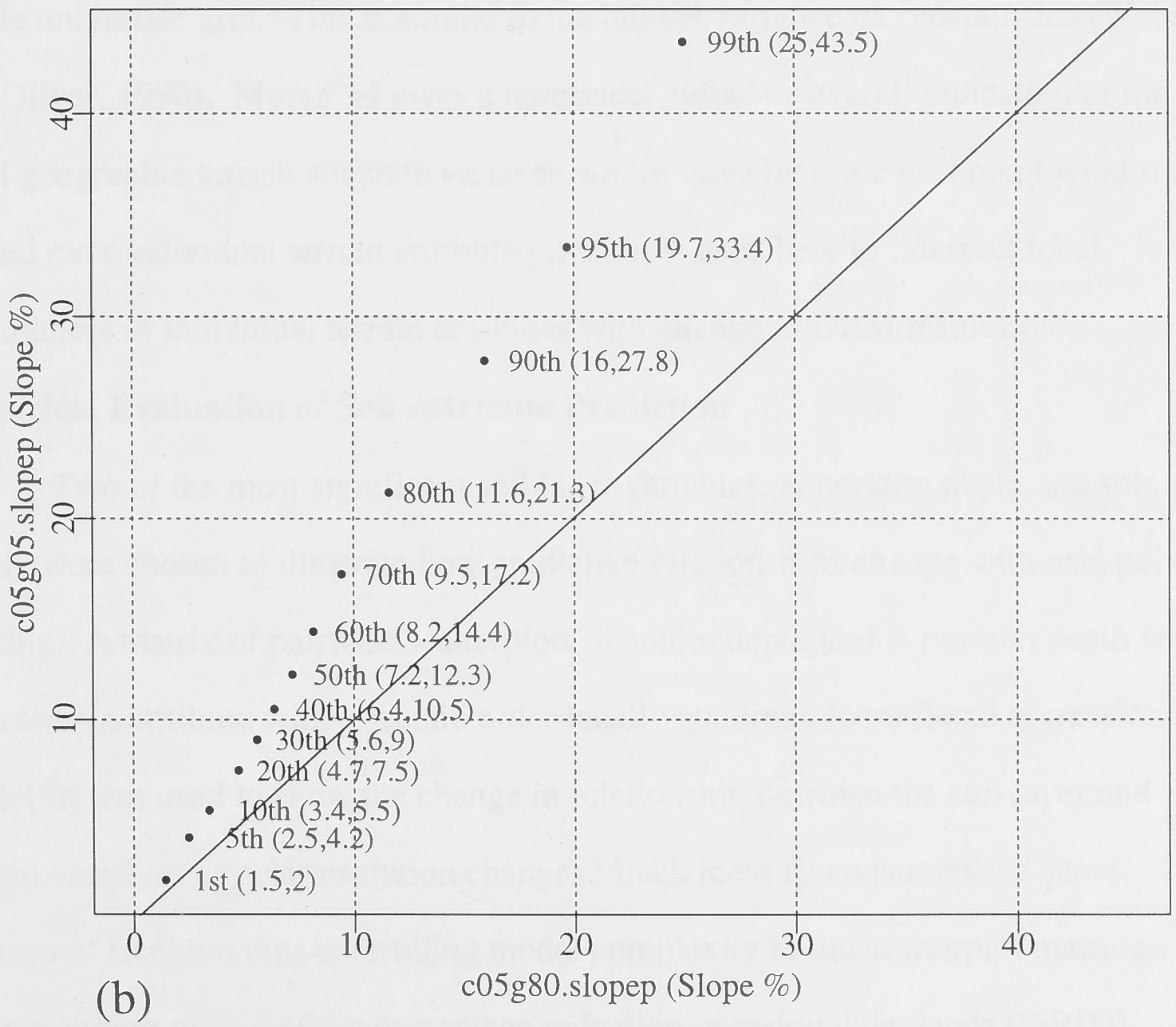
Quantile-quantile or Q-Q plots (Wilk and Gnanadesikan, 1968; Cleveland, 1993) were used to compare terrain attribute distributions. As previously discussed (Section 2.4.1), the f quantile, q(f), of a set of data is a value along the measurement scale of the data with the property that a fraction f of the data are less than or equal to q(f). For example, q(0.5) is the median value of a distribution where half of the data in the distribution are less than this value. The f-values provide a standard for comparison independent of the total number of samples in a univariate distribution. Thus, the f-values for data derived from DEM's with different resolutions and numbers of cells can be compared.

The thirteen f-values used here for comparison of terrain attribute distributions are 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99 and range from the 1st percentile to the 99th percentile of the distribution. The f-values were selected to provide an overall characterization of a distribution with additional detail in the tails. A matrix of pairwise Q-Q plots was generated for comparing attribute distributions over the range of scales (5-80m) generated from the two topographic data sources.

Figure 4.4a illustrates the pdf's of the distributions for slope percent computed from the five metre contour source data (c05) deriving a 80m grid (g80) versus the same source data deriving a 5m grid (g05). Figure 4.4b shows the corresponding Q-Q plot for the same data. The solid diagonal line of Figure 4.4b indicates where the quantiles should fall if the distributions are equivalent. For any quantile q(f), if the abscissa (x) is greater than the ordinate (y), q(f) will plot to the



(a)



(b)

Figure 4.4 Slope Distributions (a) and Q-Q Plot (b)

right of the solid line and vice versa. Dashed lines representing intervals in slope attribute space (e.g. 0, 10, 20, 30, 40% slope) are added for reference.

In figure 4.4b, the quantiles plotting left indicate that slopes computed from the 80m grid are smaller than those computed from the 5m grid. The c05g80.slope first quantile is seventy-two percent of the c05g05.slope first quantile (1.5, 2) and decreases to fifty-seven percent at the ninety-ninth quantile (25, 43.5). The largest decrease occurs at steeper slopes, but an almost linear curve can be drawn through the q(*f*) points suggesting that the distribution change with scale is systematic. Matrices of Q-Q plots can be used to efficiently compare large numbers of distributions to synthesize an understanding of change with scale. Further interpretations are discussed below.

Complementary information to the univariate distribution comparisons is provided by Moran's *i* statistic (Goodchild, 1986) which measures the strength of autocorrelation or the similarity of grid point values adjacent to one another in a single univariate grid. This is similar to the nugget variance of geostatistics (Webster and Oliver, 1990). Moran's *i* gives a numerical global or overall indication of the local geographic terrain attribute variance within three by three grid point windows passed over individual terrain attribute grids. It is used here to interpret local smoothness of individual terrain attributes with change in DEM resolution.

Empirical Evaluation of Soil Attribute Prediction

Two of the most significant soil layer variables, A horizon depth and solum depth, were chosen to illustrate how predictive relationships change with grid point spacing. A matrix of pairwise scatterplots of solum depth and A horizon depth versus each terrain attribute with a second order locally quadratic loess (local regression) model fit was used to show the change in relationship between the soil layer and terrain variables as grid resolution changes. Each loess fit consumes the same degrees of freedom thus controlling model complexity in the scatterplot matrices. A corresponding matrix of the percentage reduction in residual deviance (%RID)

provides a quantitative measure of the variation accounted for by the loess model illustrated in each scatterplot.

4.3 RESULTS & DISCUSSION

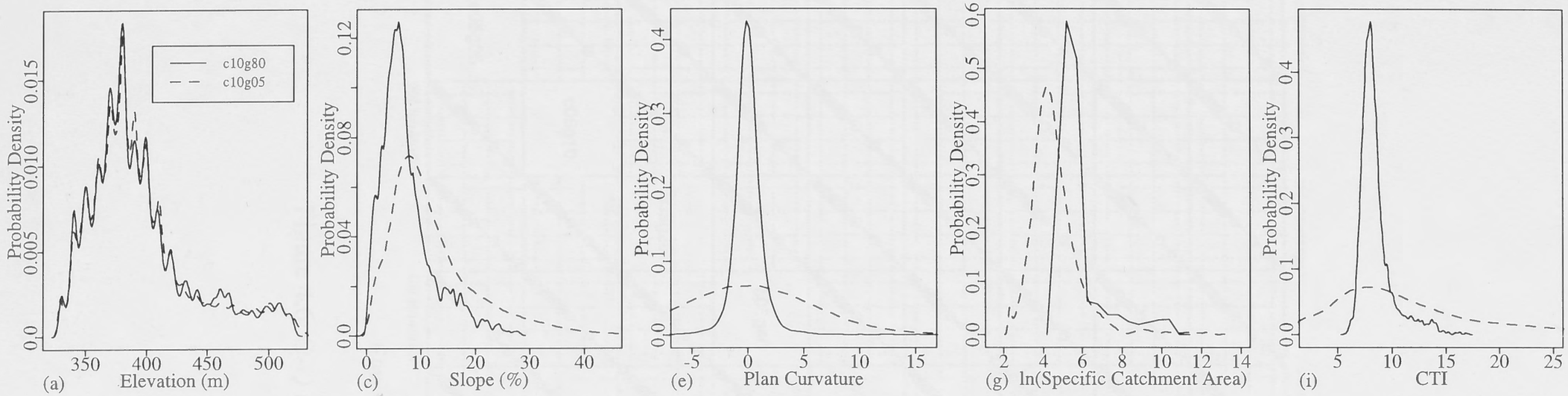
4.3.1 DEM Resolution

Figure 4.5 shows the probability distribution functions for the smallest and largest grid point spacings from both data sources for elevation, slope, natural log of specific catchment area, plan curvature and the compound topographic index. Figures 4.6 to 4.10 show the corresponding pairwise Q-Q plots of these terrain attributes generated from the different grid spacings and source data DEM's. These provide a comparison of a range of terrain variables namely the primary data (elevation), primary attributes (slope, plan curvature) and secondary or compound attributes (specific catchment area, compound topographic index). The individual matrices of Q-Q plots (Figures 4.6-4.10) are arranged as a mirror of Q-Q plots with those plots above the diagonal being a replicate of those below the diagonal with the corresponding data source and grid resolution (e.g. c05g05 & c05g80) swapping x- and y-axes. The source data and grid spacing corresponding to the x- or y-axis of any individual Q-Q plot is defined by the abbreviated names located on the plot diagonal. Dotted lines are placed on each individual Q-Q plot for reference.

Elevation

Figures 4.5a, 4.5b, and 4.6 display the corresponding pdf's and Q-Q plots for elevation. These show almost no change in distribution for the primary elevation DEM between scales and data sources. Figures 4.5a and 4.5b suggest that the more detailed data source (c05) provides a smoother density function exhibiting fewer peaks and valleys. ANUDEM is known to exhibit a terracing affect on areas away from the contours (Hutchinson, personal communication) or data sparse regions. The lack of prominent peaks and valleys in DEM's generated from the more detailed data suggests that the extra information is providing a better representation in these areas.

<-- top row - c10g10 (dashed line) and c10g80 distributions (solid line) -->



<-- bottom row - c05g05 (dashed line) and c05g80 distributions (solid line) -->

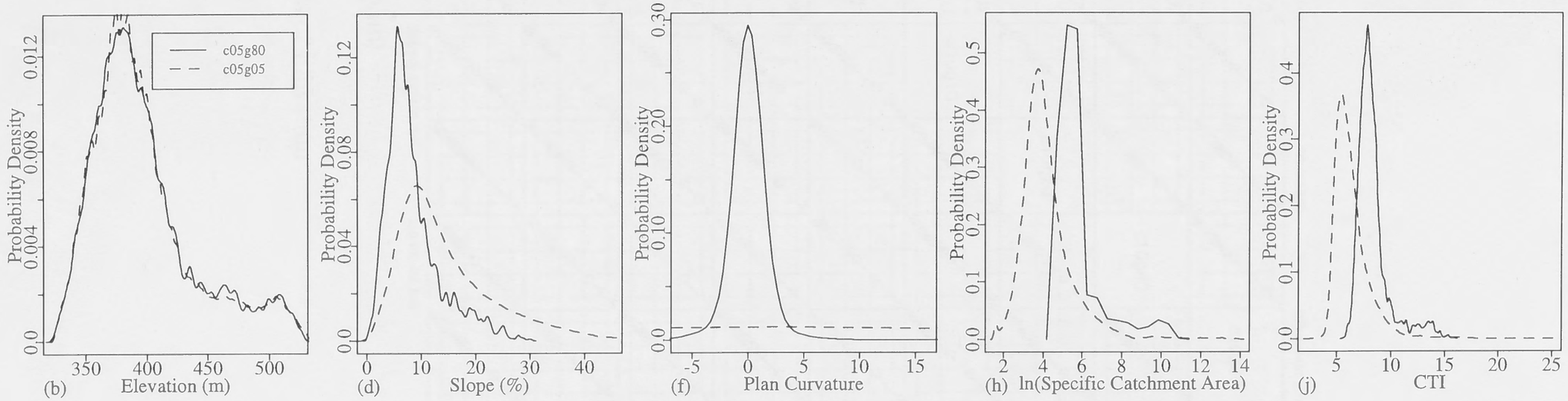


Figure 4.5(a-j) Varied Resoulution Terrain Attribute PDF's

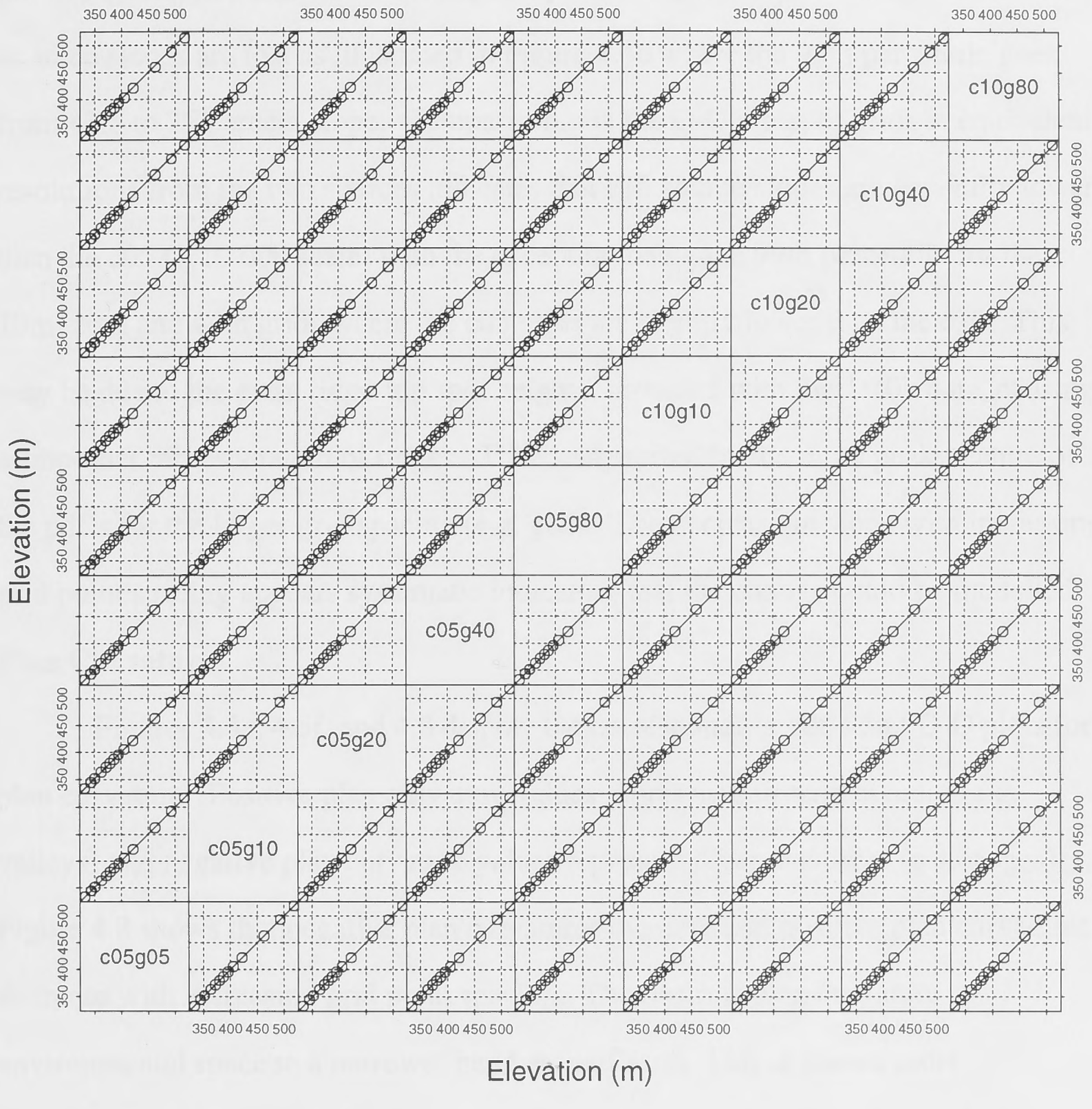


Figure 4.6 Q-Q Plots of Elevation

Slope

Figures 4.5c, 4.5d, and 4.7 display the corresponding pdf's and Q-Q plots for slope. The Q-Q plots show that slope decreases steadily with increasing grid point spacing. Steep slopes show the greatest difference as shown by the overall shift in pdf in Figures 4.5c and 4.5d. Consequently, the range in slope decreases dramatically as steep slopes are lost as illustrated in Figure 4.4b where the 99th percentile goes from 43.5 to 25 percent slope. Comparison, via the Q-Q plots, of grids at equivalent resolutions from the two sources indicates that c10 source slopes are generally lower than the c05 (1:10 000 scale) with the exception being the 99th percentile for the 10m, 20m and 40m grids where the c05 grids are slightly lower than the c10. This may be due to the extra ridge-top spot heights, provided with the 1:10k data, causing a smoother representation of ridges. This is supported by the more peaky nature of the pdf's for the larger grid spacing (e.g. g80). The decrease in slope with increasing grid point spacing appears systematic indicating that the change could be modelled.

Plan Curvature

Figures 4.5e, 4.5f, and 4.8 display the corresponding pdf's and Q-Q plots for plan curvature. Positive plan curvature values represent convergent areas (e.g. valleys) and negative plan curvature values represent divergent areas (e.g. ridges). Figure 4.8 shows that negative plan curvatures increase and positive plan curvatures decrease with increasing grid point spacing. The restricts plan curvature environmental space to a narrower band around zero. This is shown more dramatically in Figures 4.5e and 4.5f where plan curvature computed at the detailed grid spacings shows a much wider distribution. The Q-Q plots suggest that changes are greatest in the distribution tails. These influences are likely the result of a smoothing of the landscape with increasing grid point spacing as visualized in the hillshades of Figures 4.2 and 4.3.

Comparison of equivalent grid spacings from different data sources shows that the 1st percentile is always larger (less negative) for the c10 distributions. For the

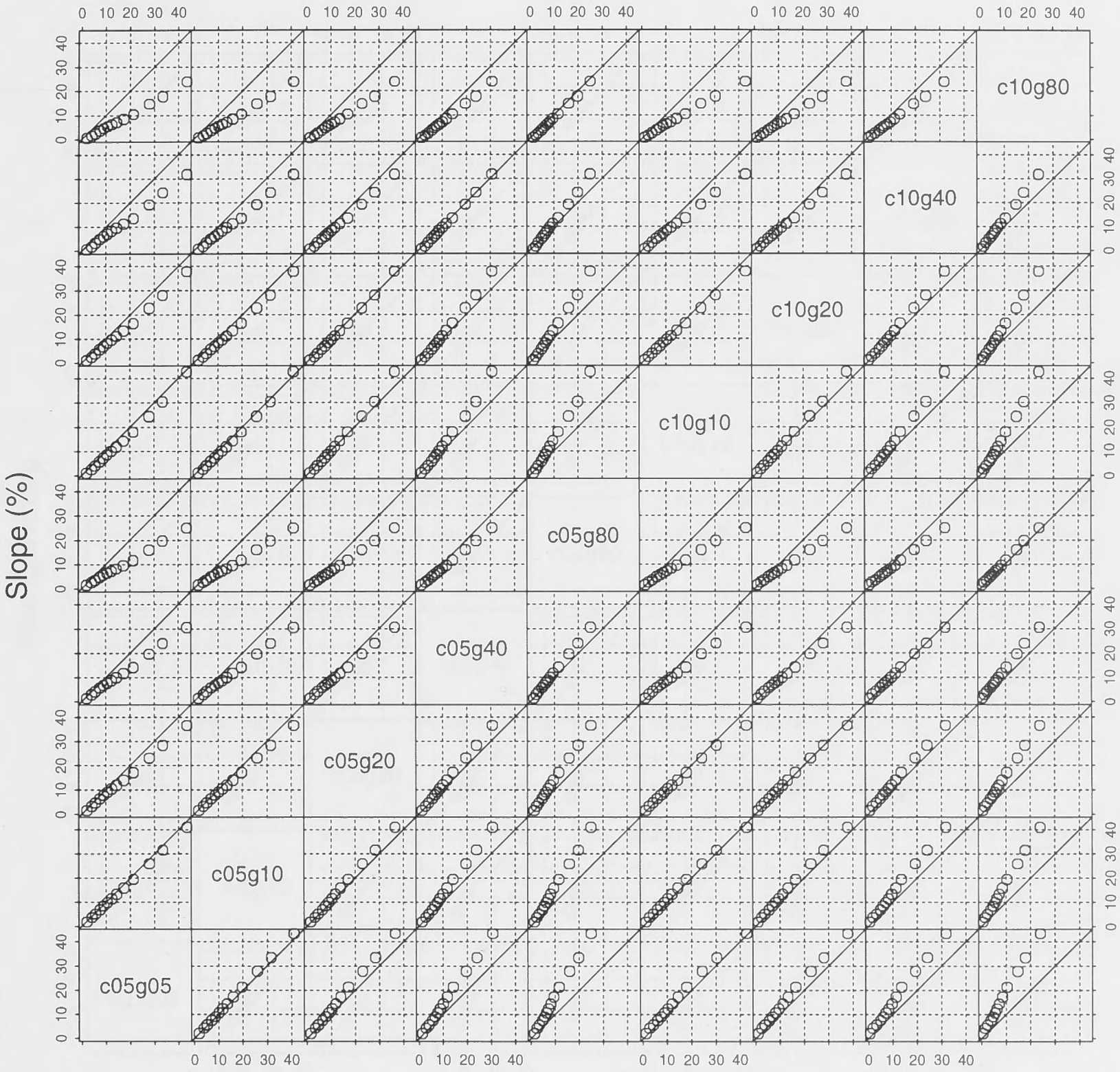


Figure 4.7 Q-Q Plots of Slope

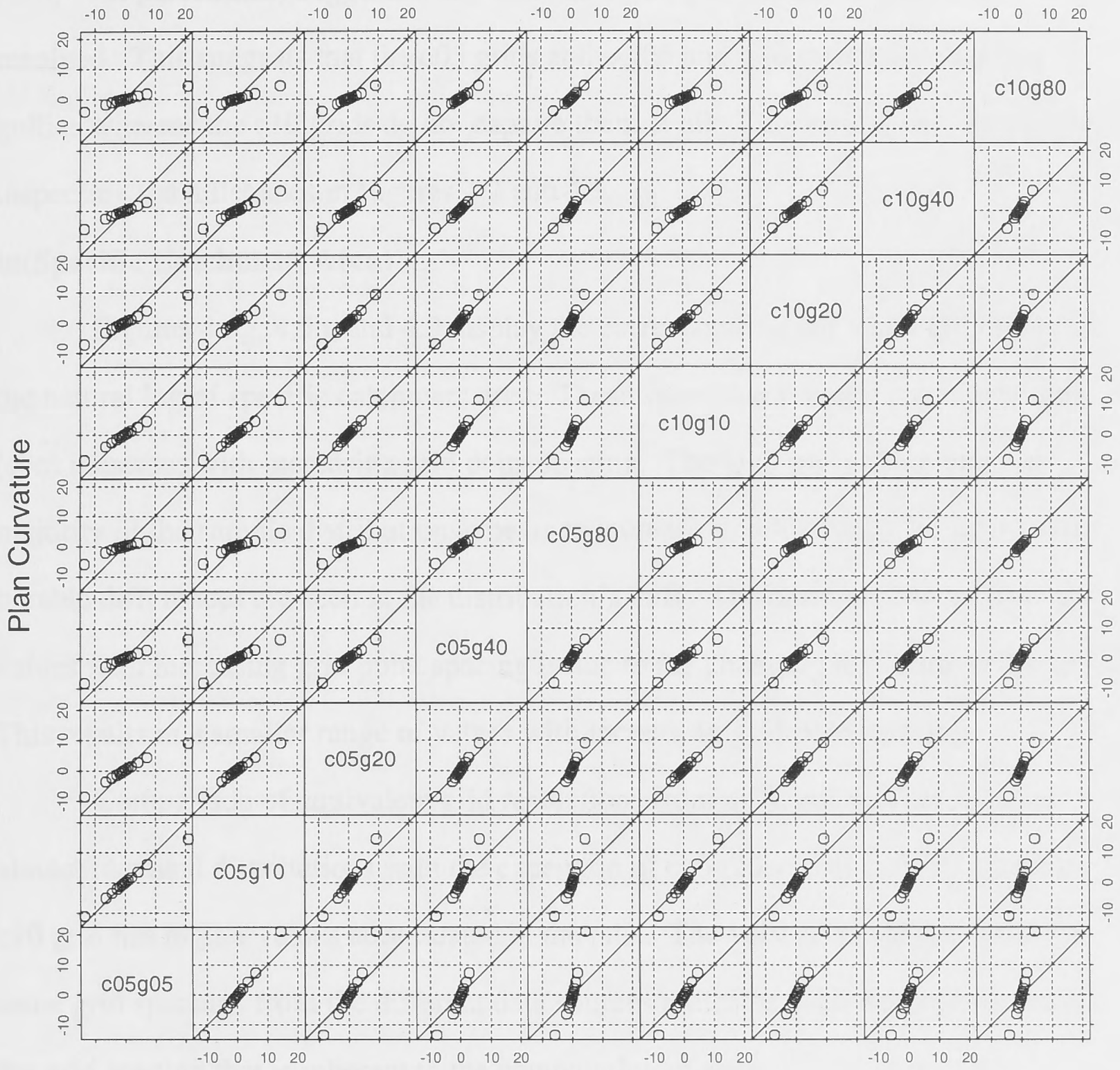


Figure 4.8 Q-Q Plots of Plan Curvature

g20 through g80 sizes the rest of the distribution is nearly equivalent. The c05g10 versus c10g10 is the exception showing that the c10 distribution is pulled inwards towards zero. The difference in the 1st percentile likely relates to the extra point data on ridge-tops for the c05 grids. The difference in the positive plan curvature tail (e.g. 95th, 99th percentiles) suggests a key scale at which gullies can no longer be resolved. This suggests that the c05 grids at the g05 and g10 resolutions capture gullies whereas the c10 grids do not capture them at all. This can be seen by carefully inspecting the hillshades in Figures 4.2 and 4.3.

ln(Specific Catchment Area)

Figures 4.5g, 4.5h, and 4.9 display the corresponding pdf's and Q-Q plots for the natural log of specific catchment area. These show that specific catchment area (spc) increases with increasing grid point spacing. The Q-Q plots show that the majority of the sample distribution appears to increase steadily with size while more notable differences are seen in the distribution's tails. The dramatic loss of low spc values with increasing grid point spacing is due to the changing minimum cell size. This results in a smaller range of values with increasing grid point spacing.

Comparison of equivalent grid resolutions from different sources indicate almost identical distributions with the exception of c05g20 versus c10g20 where the c10 grid has higher values accentuated in the tails. The very good agreement at the same grid spacings from the different data sources indicates a strong dependence on the grid spacing that is inherent to the computation of spc.

Q-Q plot comparisons of g05 or g10 distributions with the g40 and g80 distributions shows the larger grid point spacings have much larger spc's at the 95th percentile. This may relate to a larger percentage of grid cells containing large catchment areas normally associated with stream lines. Accumulated catchment areas are coerced into large cell sizes well beyond the width of a normal stream and result in a large proportion of the distribution having large catchment areas. There is no

shift at the 99th percentiles because, at that stage, all drainage has converged to a few major streamlines.

Compound Topographic Index

Figures 4.5i, 4.5j, and 4.10 display the corresponding pdf's and Q-Q plots for the compound topographic index. These indicate that CTI increases with increasing grid point spacing suggesting that the increase in specific catchment area with grid spacing is over-riding the decrease in slope with point spacing. The more extreme departure at the 95th percentile between the small and large grid point spacings, as evident in the spc comparisons, provides support to this interpretation. Comparison of equivalent grid point resolutions from the different sources show the c10 sources generate a slightly larger CTI most prominent at the g10 and g20 point spacings. This is reflected more dramatically in a comparison of the c10g10 and c05g05 distributions shown in Figures 4.5i and 4.5j. The c10g10 CTI distribution shows a larger proportion of the distribution at higher CTI's. The likely explanation is that the lower slopes for the steep areas of the landscape computed using the c10 source data, as mentioned above in discussion of slope, are causing a shift in the CTI distribution.

Spatial Autocorrelation

Tables 4.1 and 4.2 show the Moran's *i* coefficient for all grids. In general, the strength of local spatial autocorrelation decreases (i.e. Moran's *i* decreases) with the order of DEM derivative (e.g. elevation, slope, curvatures). Specific catchment area exhibits relatively low autocorrelation overall. Profile curvature shows much stronger autocorrelation than plan curvature indicating that the rate of change of slope (profile curvature) has a lower spatial frequency or is more smoothly varying than local flow convergence and divergence (plan curvature). CTI exhibits moderately strong autocorrelation that decreases steadily with increasing grid point spacing indicating less spatial coherence or connectivity at larger point spacings. The upslope mean plan and profile curvature grids follow the patterns exhibited by their respective local curvature measures.

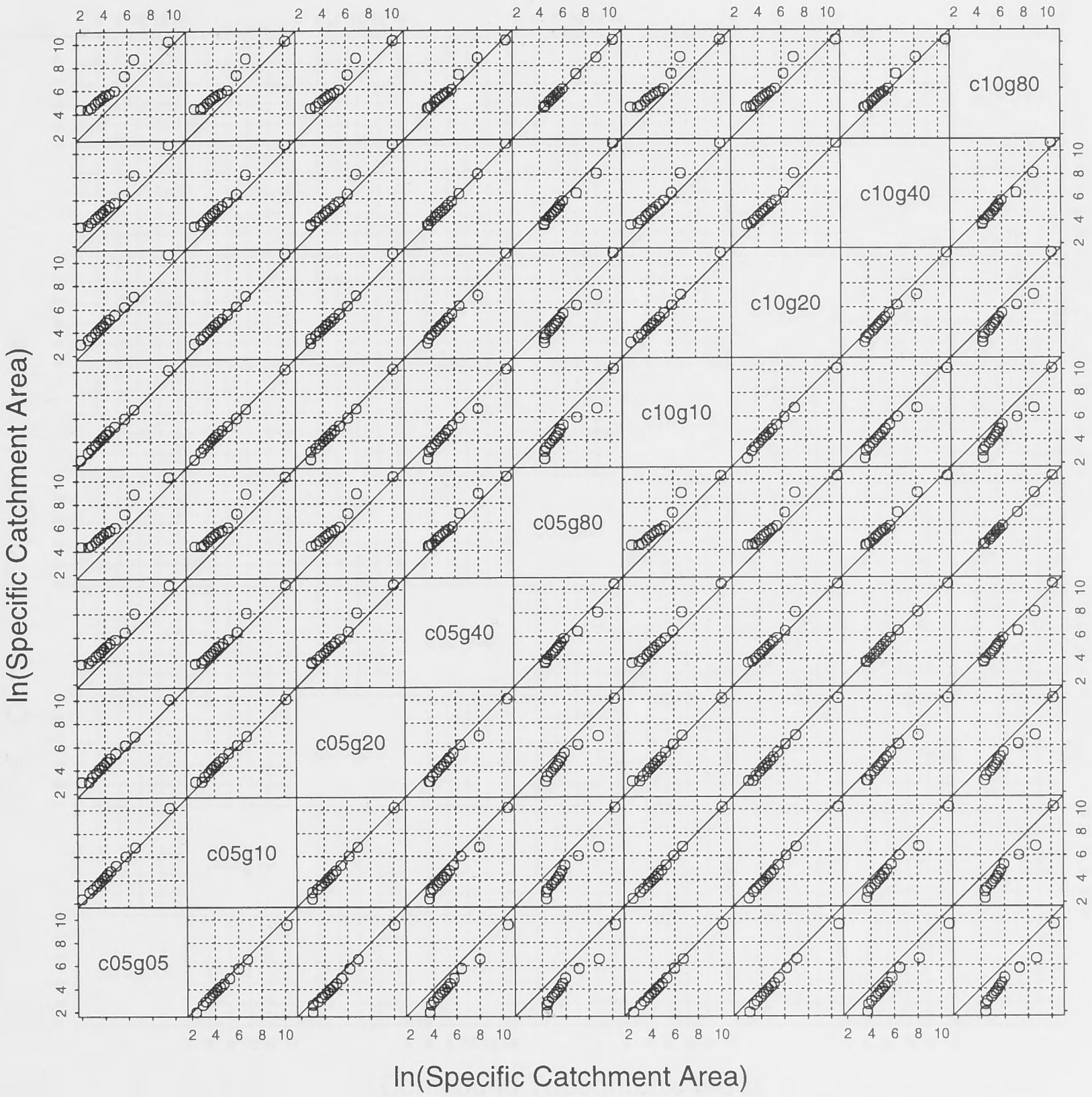


Figure 4.9 Q-Q Plots of ln(Specific Catchment Area)

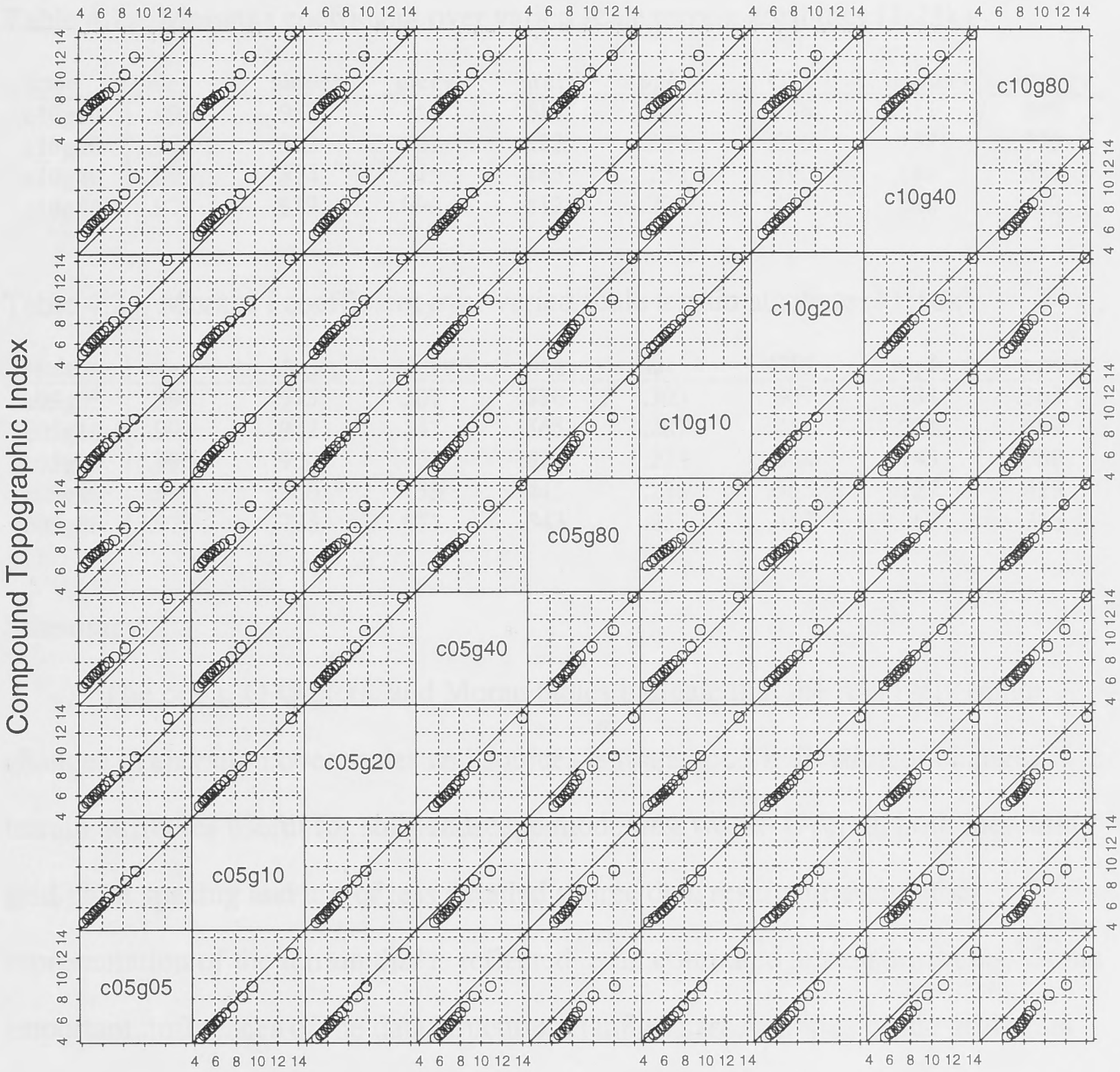


Figure 4.10 Q-Q Plots of CTI

Comparison of Moran's *i* at equivalent grid point spacings between sources shows that the c10 grids are consistently higher indicating stronger spatial autocorrelation in the c10 grids likely relating to the smoother representation of the terrain from the 1:25 000 source data.

Table 4.1 Morans *i* coefficient over varied scale terrain attributes (1:25k)

scale	**elev**	**slopep**	**plcrv**	**prcrv**	**spc**	**CTI**	**mplcrv**	mprcrv
c10g10	.998	.957	.356	.613	.284	.786	.311	.599
c10g20	.995	.923	.160	.522	.280	.686	.157	.579
c10g40	.990	.874	.215	.485	.254	.590	.188	.537
c10g80	.979	.839	.096	.412	.275	.484	.080	.452

Table 4.2 Morans *i* coefficient over varied scale terrain attributes (1:10k)

scale	**elev**	**slopep**	**plcrv**	**prcrv**	**spc**	**CTI**	**mplcrv**	mprcrv
c05g05	.999	.973	.207	.586	.303	.809	.184	.565
c05g10	.998	.952	.243	.565	.287	.752	.233	.589
c05g20	.995	.920	.162	.571	.238	.668	.143	.580
c05g40	.990	.860	.153	.445	.238	.585	.127	.429
c05g80	.977	.775	.071	.343	.252	.457	.067	.328

Summary

The pdf's, Q-Q plots, and Moran tables indicate that the relatively minor changes in elevation over scales and sources is not a good indication of changes in the terrain attributes useful for soil-landscape modelling work. Overall, both increasing grid point spacing and use of less detailed source data results in a smoother representation of the terrain that is reflected in all computed terrain attributes. Some important influences of the data structure relating to assumptions of the grid point applying to a grid cell area were discovered for specific catchment area. Large catchment areas normally associated with stream channels are overly represented at large grid point spacings (e.g. 40m, 80m). Likewise, small catchment areas close to ridge-tops and drainage divides are lost because they can not be represented by large grid spacings. Specific catchment area is an important variable for hydrology. Scale related artefacts are therefore easily propagated in indices such as CTI.

With the exception of specific catchment area and related secondary attributes (e.g. CTI), the terrain attribute distribution changes relating to grid point spacing appear systematic indicating that they could be modelled. Although patterns exist in Moran's *i* coefficient (e.g. continuously decreasing over scales), these are not as systematic and indicate that it would be difficult to model spatial autocorrelation changes with scale.

4.3.2 Soil Attribute Prediction

Figures 4.11 to 4.14 show the scatterplot matrices of soil attributes versus explanatory terrain attributes from the different DEMs for the study area. The matrix columns are arranged according to explanatory terrain attribute with increasing grid point spacing by row. The solid line represents a locally quadratic regression (loess smoother) fit. Tables 4.3 to 4.6 show the percentage reduction in deviance for each of the loess model fits. Percentage reduction in deviances greater than fifty percent are indicated by a star next to the reduction in deviance value.

A Horizon Depth

Figures 4.11 and 4.12 and Tables 4.3 and 4.4 illustrate and quantify relationships between A horizon depth and terrain attributes. Elevation provides little predictive utility, but the scatterplot and predictive pattern as illustrated by the loess fit remains stable across scales and DEM source. The approximate horizontal line for slope percentage over the c10 scales indicates a mostly random scatter exhibiting little predictive utility for A horizon depth. This is echoed in Table 4.3. The c05 loess fits show a general monotonic increase with slope percentage that degrades as grid size increases, but the predictive utility is low.

The natural log of specific catchment area proves highly significant over several scales from both data sources. The c10g10 loess fit provides the largest A horizon depth %RID (61.60) of any explanatory terrain attribute. The general pattern of the line, monotonically decreasing with increasing catchment area, maintains a similar negative slope over the c10 scatterplots. A comparable loess fit slope is

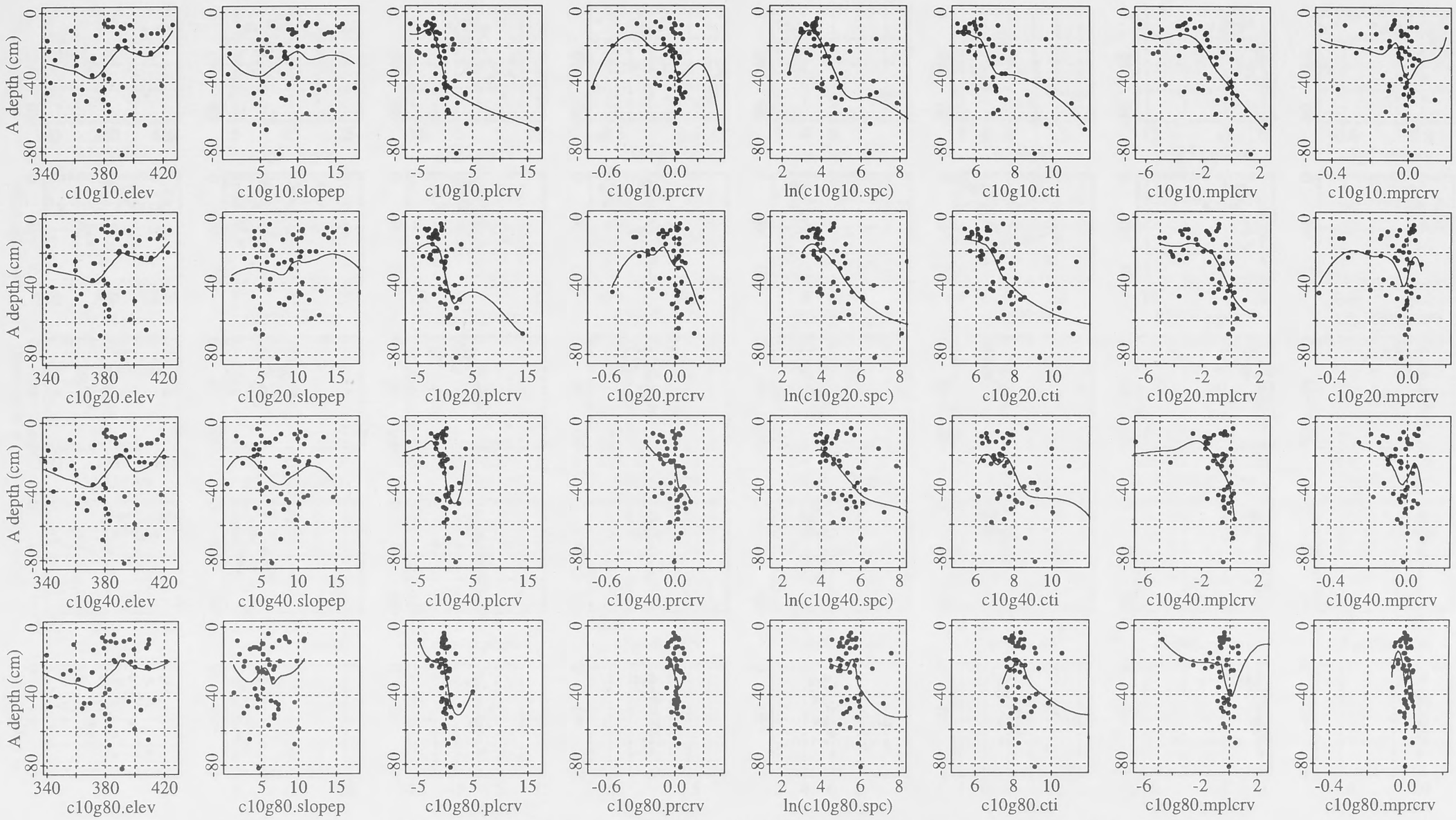


Figure 4.11 A Horizon Depth versus Terrain Attributes (1:25 000 source data - Griggward)

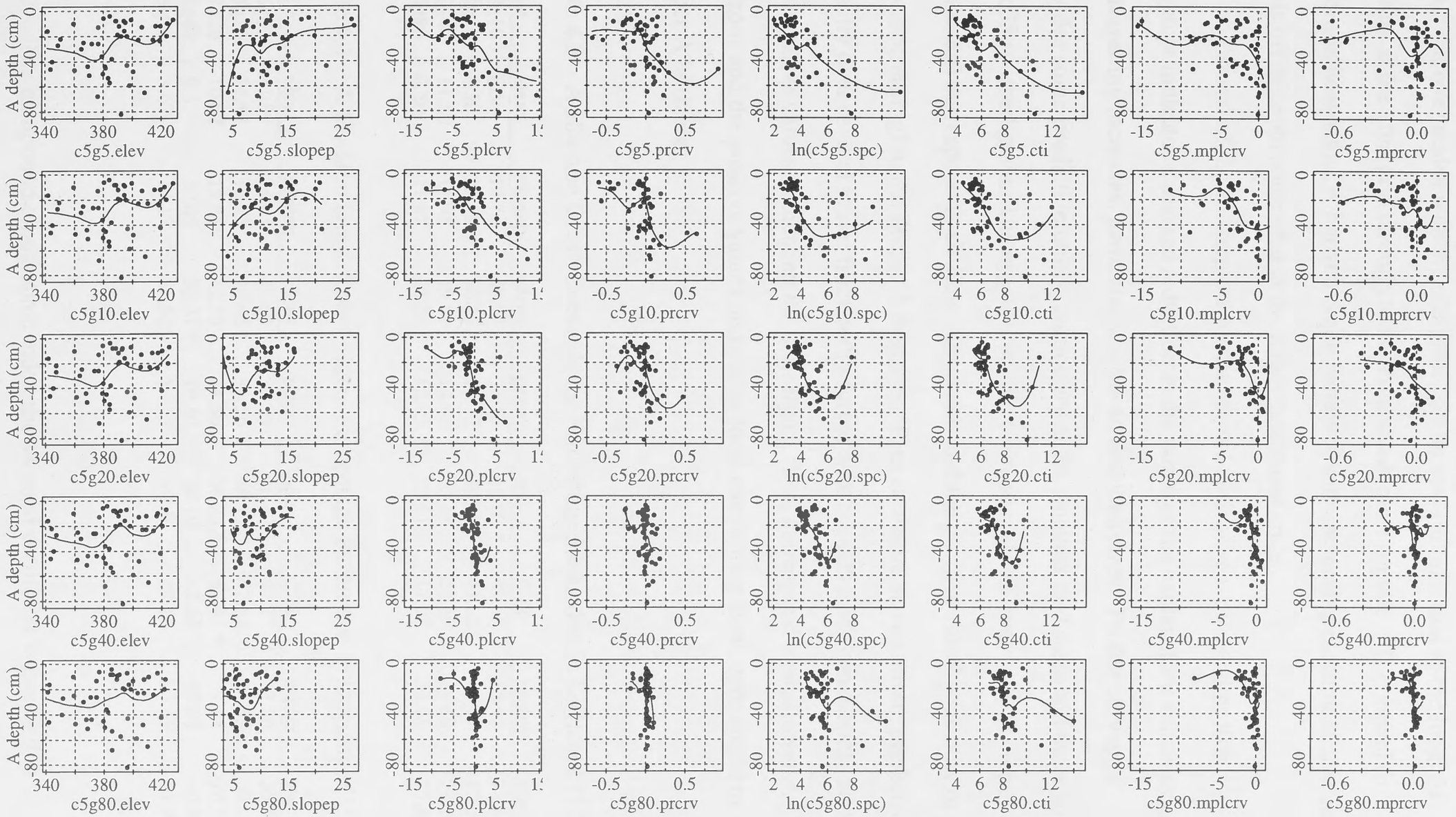


Figure 4.12 A Horizon Depth versus Terrain Attributes (1:10 000 source data - Griggward)

exhibited over scales with the c05 grids, but an upward inflexion in the loess fit appears at the 10m grid through to the 80m scatterplots for c05. The model fit pattern indicates that A horizon depth increases with increasing catchment area and levels off in depth somewhat at the larger catchment areas.

Perusal of the scatterplots for the second derivatives, plan and profile curvature indicates a marked reduction in the range of the explanatory values as grid point spacing increases, particularly at the 40 and 80m sizes. Profile curvature provides some predictive utility, particulary at the smallest grid sizes and more so for the c05 sourced grids. The negative profile curvatures (local convexity or slope increasing) correspond to shallow A horizons and the positive (local concavity or slope decreasing) to the deeper A horizons. Plan curvature shows strong predictive capacity over several scales from both sources with the c05 grids consistently better. The negative values, indicating local diverging flow, correspond to shallower A horizons and the positive values, indicating local converging flow, correspond to deeper A horizons.

Table 4.3 A Horizon Depth Loess Model Percentage Reduction in Deviance (1:25k)

scale	**elev**	**slopep**	**plcrv**	**prcrv**	**spc**	**CTI**	**mplcrv**	mprcrv
c10g10	8.56	13.82	46.36	35.38	61.60 *	49.65	44.90	27.44
c10g20	8.50	15.89	52.36 *	22.00	49.83	47.62	41.39	17.04
c10g40	11.80	11.95	39.58	14.48	33.89	28.69	26.80	20.24
c10g80	0.56	11.87	22.26	12.93	23.51	17.07	23.58	27.43

Table 4.4 A Horizon Depth Loess Model Percentage Reduction in Deviance (1:10k)

scale	**elev**	**slopep**	**plcrv**	**prcrv**	**spc**	**CTI**	**mplcrv**	mprcrv
c05g05	11.99	24.63	42.94	33.55	46.37	47.37	23.50	35.49
c05g10	11.62	17.43	50.55 *	38.26	46.12	42.57	44.53	30.68
c05g20	10.02	19.88	42.55	31.36	56.83 *	57.82 *	37.18	20.53
c05g40	9.37	20.05	55.31 *	19.10	39.10	42.46	49.65	40.34
c05g80	3.67	17.17	36.52	14.75	-0.40	27.18	25.99	21.91

The compound topographic index shows good predictive capacity that degrades sharply at the g80 point spacings. The loess fit pattern closely resembles

that of the component specific catchment area indicating strong collinearity between these explanatory variables. The contextual upslope mean plan and profile curvatures show useful predictive patterns, with plan curvature showing greater predictive capacity at the c10g10, c10g20 and more broadly over the c05 grid sizes.

In summary, no individual grid point spacing stands out as best for modelling A horizon depth. Instead several grid point spacings exhibit useful predictive capacity. Specific catchment area and plan curvature are the most useful terrain attributes showing strong predictive capacity over several scales from both topographic data sources. These two attributes characterize landscape processes at fundamentally different scales, with specific catchment area indicating the hydrological and geomorphic catchment context and plan curvature the local landform concavity and convexity relating to local flow convergence and divergence. Slope gradient does not appear as useful in this study area. Relationships were the least useful at the 80m grid point spacing. The c10 source grids decreased in predictive utility with increase in point spacing while the c05 grids appear to maintain predictive utility over more scales. The c10g20 attributes, overall, exhibit less predictive capacity than the c10g10 predictors but are better than the c10g40 and c10g80 attributes.

Because the A horizon is the surface layer, it is more susceptible to land management influences and ephemeral events that can make it difficult to model spatially. Therefore, the presence of useful landform predictive relationships over several scales in this study area is very encouraging and suggests that these attributes are characterizing patterns that relate to significant landscape processes affecting A horizon formation.

Solum Depth

Figures 4.13 and 4.14 and Tables 4.5 and 4.6 illustrate and quantify relationships between solum depth and terrain attributes. Elevation is of little predictive use, but again maintains a similar pattern over scales and source. Slope is

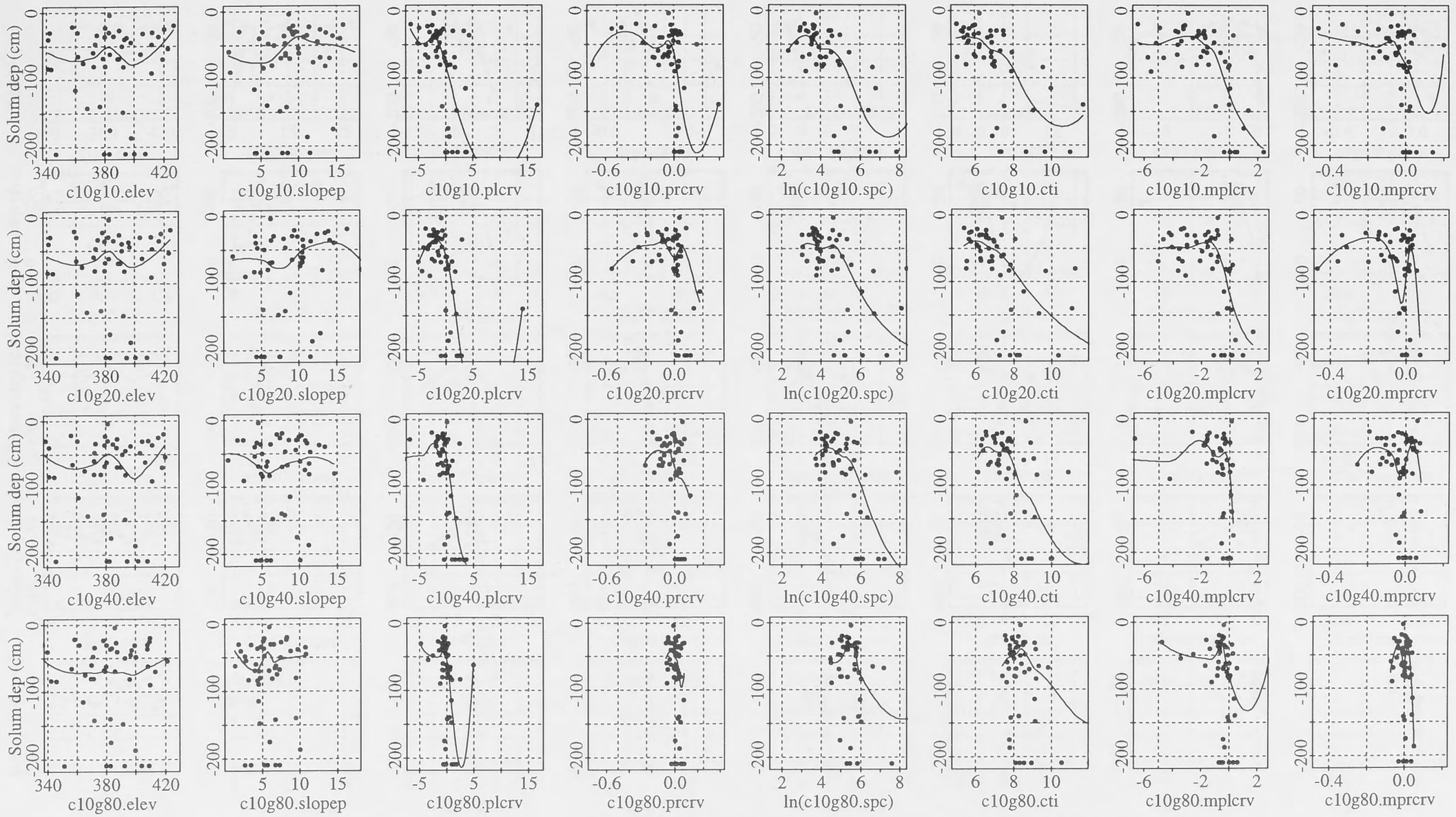


Figure 4.13 Solum Depth versus Terrain Attributes (1:25 000 source data - Griggward)

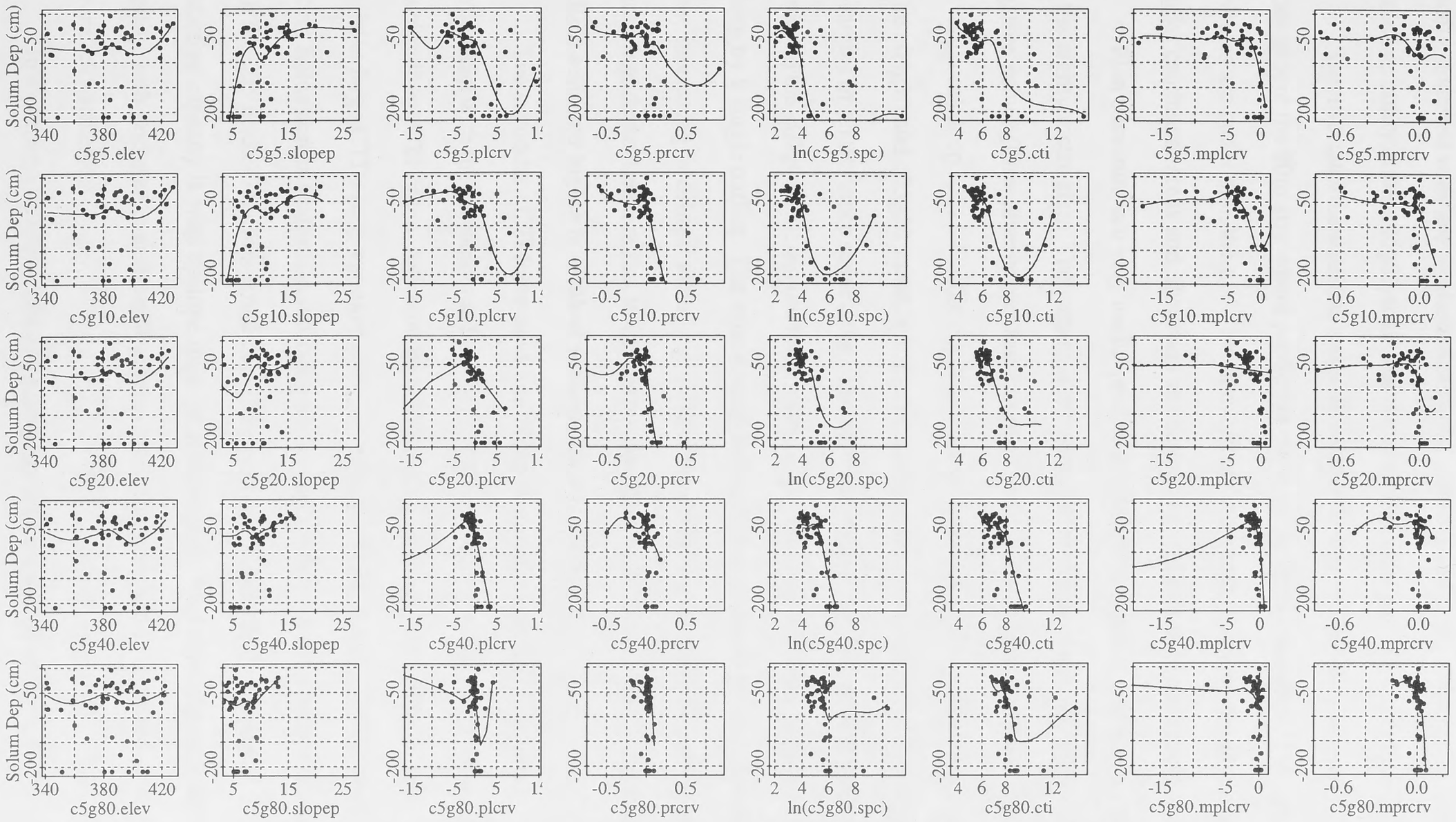


Figure 4.14 Solum Depth versus Terrain Attributes (1:10 000 source data - Griggward)

broadly scattered with low predictive potential over the c10 grids, but exhibits strong predictive utility at the c05g05 scale and declines with grid size for the c05 grids.

Specific catchment area is a significant explanatory variable over several scales except the 80m size where predictive capacity declines markedly. The c05 spc attributes are better than the c10. The loess fit and scatterplot pattern is tightest at the smaller catchment areas and illustrates an increase in variance with catchment area.

Plan curvature is a very useful predictor across data sources and scales except for the eighty metre size. The pattern is the same as for A horizon depth where the negative values (flow divergence) indicate shallow soils and positive values (flow convergence) deep soils. Profile curvature is not generally useful with the exception of the c05g10 and c05g20 variables. The c05g20 predictors exhibit a 72.37 %RID, the largest of all deviance reductions. The scatterplots indicate that the pattern is present in the c05g05 plot but the loess model is pulled away from the general scatter pattern by a single outlier. The relationship indicates that negative profile curvatures (slope increasing, erosional areas) correlate to the shallow soils and positive (slope decreasing, depositional areas) to the deep soils. The model fit is, again, better for the shallow soils and higher in residual variance for the deeper soils.

The secondary CTI loess model exhibits an initially tight (low variance) relationship with the predictors at low CTI's followed by a consistent negative slope down to about CTI of ten, followed by a leveling off and broader scatter of soil depths at large CTI's (e.g. variance increases with the fitted mean). This general pattern holds across scales and sources with the largest deviance reductions occurring with the c05 grids. The contextual curvature predictors indicate the strongest predictive capacity is with upslope mean plan curvature. Predictive capacity declines steadily with grid size over the c10 sources while the potential is scattered over several c05 scales.

In summary, no individual c10 grid point spacing stands out best for modelling solum depth, but g80 is definitely the worst. The c10 predictors decline in

predictive capacity with increasing grid spacing and show that specific catchment area, CTI and upslope mean plan curvature show the strongest predictive capacity. Although the c10g20 attributes show good predictive relationships for several attributes, the best relationships are at the c10g10 point spacing.

Table 4.5 Solum Depth Loess Model Percentage Reduction in Deviance (1:25k)

scale	**elev**	**slopep**	**spc**	**plcrv**	**prcrv**	**CTI**	**mplcrv**	**mprcrv**
c10g10	10.86	15.82	57.06 *	42.81	27.46	50.93 *	53.34 *	30.01
c10g20	11.27	19.00	47.33	44.84	29.60	46.53	45.75	38.36
c10g40	11.40	7.85	44.65	40.95	15.64	35.49	18.06	16.32
c10g80	2.08	6.16	15.15	27.51	13.16	18.73	16.61	22.03

Table 4.6 Solum Depth Loess Model Percentage Reduction in Deviance (1:10k)

scale	**elev**	**slopep**	**spc**	**plcrv**	**prcrv**	**CTI**	**mplcrv**	mprcrv
c05g05	10.28	51.46 *	52.54 *	46.05	22.17	52.45 *	30.31	27.18
c05g10	9.70	32.11	65.35 *	45.99	63.61 *	59.87 *	50.75 *	39.95
c05g20	9.11	16.56	58.46 *	58.62 *	72.37 *	57.87 *	49.42	41.36
c05g40	7.97	25.51	56.46 *	59.54 *	21.35	59.98 *	65.26 *	34.99
c05g80	6.19	17.09	6.44	28.53	19.69	31.14	30.39	25.06

While the c05 predictors are generally better than the c10 source predictors and show excellent predictive capacity for several attributes over many scales, predictions at certain point spacings stand out. The c05g20 and c05g10 profile curvature predictors show very %RID's of 72.37 and 63.61 respectively. This is peculiar because all other profile curvature point spacings show very little predictive potential. The c05g05 slope predictors also show an unusually high %RID (51.46) in comparison to all other point spacings for slope. This may indicate that the area of the 3 x 3 grid spacing computations is capturing important local processes over slope length scales of 15m and rate of change of slope (profile curvature) over the 30m to 60m length scales.

The c05 predictors show strong predictive potential for slope, specific catchment area, plan curvature, profile curvature, CTI and upslope mean plan curvature. This suggests that the c05 source data and derivatives capture key landform patterns relating to solum depth better than the c10 source derivatives. The

inclusion of attributes relating to several different landscape processes indicates that no single process dominates but that a set of processes over several scales are influencing solum depth.

4.4 CONCLUSIONS

This Chapter provides an empirical comparison of terrain attribute distributions over various grid point spacings and an evaluation of environmental correlations for spatial prediction of soil attributes. It has been conducted over a single area with a well defined physiographic domain using a fixed set of computational procedures and spatially distributed field data. The methods of analysis are useful for summarizing large quantities of data. They should be useful for comparing results from different physiographic domains.

Overall, increases in grid point spacing result in a reduction or constriction in range for terrain attribute space. The first and second derivatives (slope, plan and profile curvature) of the primary elevation data appear to change systematically with grid point spacing. Specific catchment area exhibits systematic change but includes several artefacts relating to the data structure (square cell) that complicate the changes and do not accurately characterize patterns along ridge-tops, drainage divides and streamlines at larger grid point spacings. These artefacts are propagated through to secondary attributes that incorporate specific catchment area. This work suggests that a scaling theory could be developed for moving up to larger grid point spacings in this Ordovician metasediment physiographic domain. It does not suggest how to move to more detailed scales or smaller grid point spacings from smaller scale topographic data (e.g. 1:100 000 or smaller).

The empirical findings could be used to develop a link to smaller scale representations of terrain and an understanding of how underlying probability distributions change. Although Moran's *i* coefficient provides information about the change in local spatial relations over scales, more research is needed to determine how the

distribution changes are spatially expressed. This could be built on simple representations of streamlines and ridge-tops to provide controls within which the distributions are distributed. Development of variograms to characterize spatial scales of variation for different terrain attributes at different resolutions would provide useful complementary information, but would require significant computing resources.

In general, of the grid point spacings evaluated, no individual spacing stands out as best for overall prediction of A horizon and solum depth in this study area. However, the smaller grid spacings are generally better and the 80m spacing marks a significant decline in predictive potential. Specific catchment area and plan curvature were most useful for A horizon depth prediction. Several scales and attributes were useful for solum depth prediction and indicated that solum depth was more predictable than A horizon depth. Profile curvature showed a marked increase in solum depth predictive capacity for the c05g20 and c05g10 scales. Slope gradient also stood out at the c05g05 scale. These may relate to scales that more appropriately match solum depth patterns and controlling landscape processes.

The more detailed topographic data source (1:10 000) provided better overall predictions for a range of scales indicating the extra terrain detail captured is important and that the computation techniques preserved this detail over several scales. Additional research to develop a cost-benefit analysis of the different sources would determine if the extra effort involved is worth the expenditure.

Of the c10 derived predictors, the g10 scale generally provided better predictions than the g20 suggesting that future work with the 1:25 000 topographic sources should consider interpolating to point spacings finer than the twenty metre spacing routinely used in this work. However, this will depend on the physiographic domain.

The strong predictive relationships between static quantitative landform descriptors and soil attribute patterns is very encouraging for the development of pattern/process relationships and understanding. Models were generally best in the upper catchment areas and were more scattered in lower landscape positions. This

may suggest that the upper parts of the landscape are dominated by erosional processes (i.e. losses to the system) whereas the lower parts of the landscape are a broader mixture of erosional and depositional processes (i.e. material movement, redeposition, ephemeral events etc.) causing a more complex patterning. However, this Chapter demonstrates quantitative analysis techniques from which additional hypotheses may be developed for testing.

4.5 REFERENCES CITED

Allen, T.F.H., and T.W. Hoekstra. 1992. Towards a unified ecology. Columbia University Press, New York.

Band, L.E., and I.D. Moore. 1995. Scale: landscape attributes and geographical information systems. Hydrol. Proc. 9:401-422.

Bloschl, G., and M. Sivapalan. 1995. Scale issues in hydrological modelling: a review. Hydrol. Proc. 9:251-290.

Bolstad, P.V., and T. Stowe. 1994. An evaluation of DEM accuracy: elevation, slope and aspect. Photogram. Eng. Remote Sensing. 60(11):1327-1332.

Brown, D.G., and T.J. Bara. 1994. Recognition and reduction of systematic error in elevation and derivative surfaces from 7.5-minute DEMs. Photogram. Eng. Remote Sensing. 60(2):189-194.

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, New Jersey.

Fryer, J.G., J.H. Chandler, and M.A.R. Cooper. 1994. On the accuracy of heighting from aerial photographs and maps - implications to process modellers. Earth Surface Proc. Landf. 19(6):577-583.

Gallant, J.C., 1996. TAPES terrain analysis programs. World Wide Web. http://cres.anu.edu.au/software/tapes.html.

Gallant, J.C., and M.F. Hutchinson. 1995. Scale dependence in terrain analysis. *In* Proceeding of MODSIM 95. Newcastle, NSW. 27-30 Nov. 95.

Gerrard, A.J. 1990. Soil variations on hillslopes in humid temperate climates. Geomorphology 3:225-244.

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modelling and the spatial prediction of soil attributes. Int. J. of Geog. Inf. Sys. 9(4):421-432.

Gessler, P.E., and Ashton, L.J. in prep. Wagga Wagga geographical information system database: development, structure and user access. CSIRO Divisional Working Report No.X CSIRO Division of Soils. Canberra.

Goodchild, M.F., 1986, Spatial autocorrelation. CATMOG - Concepts and techniques in modern geography. Geo Abstracts, Norwich.

Hammer, R.D., F.J. Young, N.C. Wollenhaupt, T.L. Barney, and T.W. Haithcoate. 1995. Slope class maps form soil survey and digital elevation models. Soil Sci. Soc. Am. J. 59:509-519.

Hutchinson, M.F. 1988. Calculation of hydrologically sound digital elevation models. p. 117-133. *In* Proceedings Third International Symposium on Spatial Data Handling. International Geographical Union, Sydney.

Hutchinson, M.F. 1989. A new procedure for gridding elevation and streamline data with automatic removal of spurious pits. J. Hydrol. 106: 211-232.

Hutchinson, M.F. 1995. Documentation for ANUDEM Version 4.4. Centre for Resource and Environmental Studies, Australian National University, Canberra.

Hutchinson, M.F., and T.I. Dowling. 1991. A continental hydrological assessment of a new grid based digital elevation model of Australia. Hydrol. Proc. 5:45-58.

Jenson, S.K. 1991. Applications of hydrologic information automatically extracted from digital elevation models. Hydrol. Proc. 5:31-44.

Lee, J., P.K. Snyder, and P.F. Fisher. 1992. Modeling the effect of data errors on feature extraction form digital elevation models. Photogram. Eng. Remote Sensing. 58(10):1461-1467.

McSweeney, K., P.E. Gessler, B. Slater, R.D. Hammer, J. Bell, and G.W. Petersen. 1994. Towards a new framework for modelling the soil-landscape continuum. p.127-145. *In* Factors of soil formation: a fiftieth anniversary retrospective. SSSA Special Pub. 33. Madison, WI.

Moore, I.D. 1992. Terrain analysis programs for the environmental sciences (TAPES). Agric. Sys. Inf. Tech. 4(2):37-39.

Moore, I.D., A.R. Ladson, and R. Grayson 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydro. Proc. 5:3-30.

Moore, I.D., P.E. Gessler, G.A. Neilsen, and G.A. Peterson. 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57:443-452.

Moore, I.D., A. Lewis, and J.C. Gallant. 1994. Terrain attributes: estimation methods and scale effects. p. 184-214. *In* Jakeman, Beck and McAleer (Eds.) Modelling Change in Environmental Systems. Springer, New York.

Panuska, J.C., I.D. Moore, and L.A. Kramer. 1991. Terrain analysis: integration into the agriculture nonpoint source (AGNPS) pollution model. J. Soil Water Conserv. 46:59-64.

Quinn, P., K. Beven, P. Chevallier, and O. Planchon. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrol. Proc. 5:59-79.

Speight, J.G. 1968. Parametric description of landform. p239-250. *In* G.A. Stewart (ed.) Land Evaluation. Macmillan, Melbourne.

Speight, J.G. 1974. A parametric approach to landform regions. Special Publ. no. 7. Institue of British Geographers.

Trimble Navigation Limited. 1992. GPSurvey Software Users Guide. Trimble Navigation Limited. Sunnyvale, CA.

Webster, R., and M.A. Oliver. 1990. Statistical methods in soil and land resource survey. Oxford Univ. Press, Oxford.

Wilk, M.B. and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. Biometrika 55:1-17.

Zhang, W., and D.R. Montgomery. 1994. Digital elevation model grid size, landscape representation, and hydrologic simulations. Water Resources. Res. Vol. 30. 4:1019-1028.

Chapter Five: Quantitative Soil-landscape Ecology

5.1 INTRODUCTION

5.1.1 Broad Principles

Chapter two demonstrated that many environmental variables (e.g. terrain, gamma radiometrics, climate) exhibit potential for predicting soil-landscape attributes and that a flexible exploratory data analysis and statistical modelling approach is required for modelling different soil attributes. Results also indicated that morphological soil layers are generally useful for partitioning variation for a broad range of soil attributes and that the predictive models for soil layer attributes (e.g. A horizon depth, E horizon probability, E horizon depth, solum depth) explain a large proportion of the variation in the sample sets. Chapter three demonstrated that useful terrain/soil layer attribute correlations exist at several scales or grid point resolutions and that more detailed terrain data, in general, provide better predictions. This suggests that the processes influencing soil formation and soil attribute patterns across landscapes are multi-scaled and therefore difficult to comprehensively characterize with models based on measurements taken at restricted spatio-temporal scales. This limits the overall potential for interpreting landscape processes from the collected sample evidence but provides a useful starting point for building a comprehensive understanding of pattern/process relationships.

Landscape ecology provides an appropriate set of theories and principles for interpreting soil distribution. Landscape ecology focuses on the holistic study of landscapes (Naveh and Lieberman, 1984) and quantification of the *structure*, *function* and *change* of landscapes and relationships between pattern and process (Forman and Godron, 1986; Turner and Gardner, 1991). Turner and Gardner (1991) state that *structure* refers to the spatial relationships between distinctive ecosystems, that is, the distribution of energy, materials, and species in relation to the sizes, shapes, numbers,

kinds and configurations of components. *Function* refers to the interactions between the spatial elements, that is, the flow of energy, materials, and organisms among component ecosystems. *Change* refers to alteration in the structure and function of the ecological mosaic through time.

Many of these principles are familiar to pedologists (Hole and Campbell, 1985), but we have been slow to grasp tools for a quantitative soil-landscape analysis that may, more rapidly, advance our field. Concepts of the basic structural components of the soil-landscape based on soil layers or horizons (Butler, 1959; Simonson, 1959) and the holistic study of soil-landscapes in the broader environmental context (Jenny, 1941; 1980) have existed for some time. Simonson (1959) described four broad pedogenic process groupings of: additions, losses, translocations, and transformations that, in combination or individually, cause soil horizonation. Others have contributed functional concepts for spatial hillslope hydrological and geomorphic processses (e.g. infiltration, water balance, overland flow, subsurface flow, erosion, deposition, mass wasting) that, in open drainage systems, operate within catchment or watershed systems along flow vectors or pathways (Milne, 1935; Carson and Kirkby, 1972; Ruhe, 1975; Gerrard, 1981; Selby, 1982; Walker and Butler, 1983; Buol *et al.*, 1989; Daniels and Hammer, 1992).

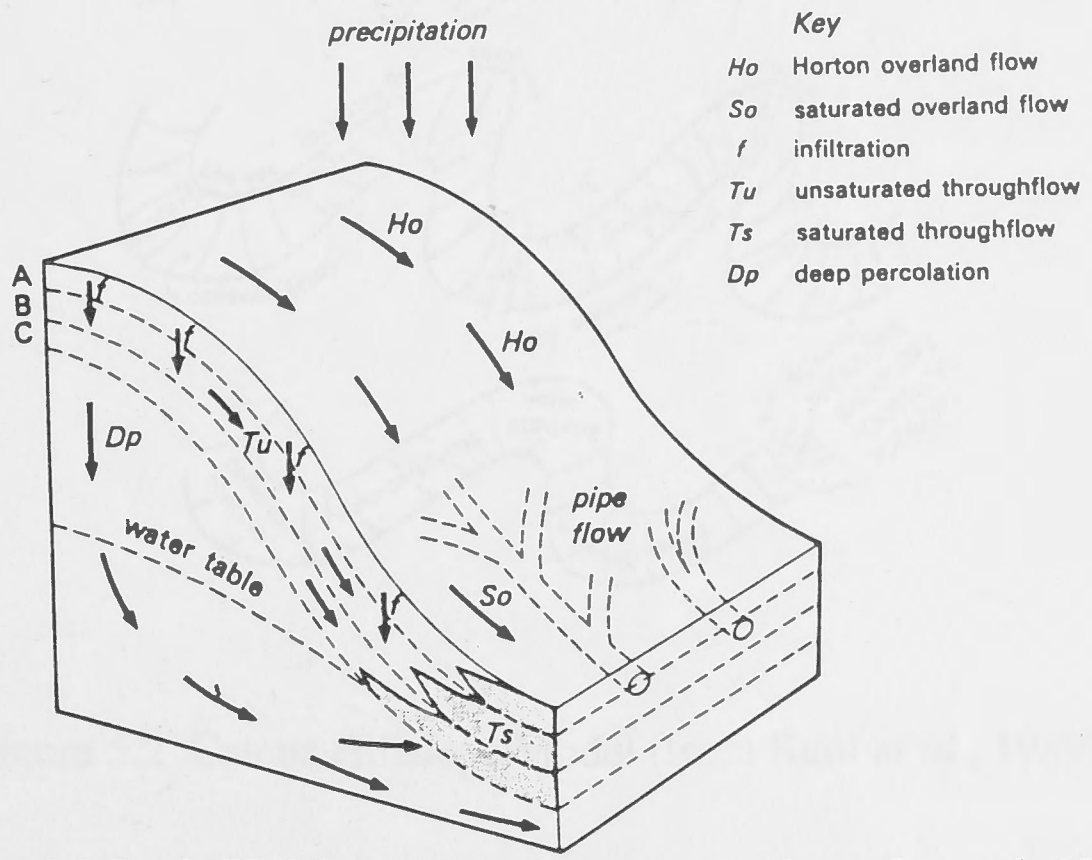


Figure 5.1 Basic Hillslope Hydrology Model (from Gerrard, 1981)

Figure 5.1 displays a simplified representation of the soil layers with planar flow vectors that provide a useful visualization of the spatial distribution and connectivity of hydrological processes operating in the landscape. This concept of connectivity includes an ordered spatial adjacency, such as hillslope summit, sideslope and base, that is important for gravitational movements of energy and material (e.g. solutes, colloids, soil particles). While these hillslope visualizations are very useful, they are rarely based on quantitative data statistically representative of a spatial area. Futhermore, in reality, we know that flow pathways converge and diverge over three-dimensional landforms (see Figure 5.2) and soil patterns often reflect these variations as a result of integrated pedogenic, hydrological and geomorphological processes (Buol *et al.* 1989). The conceptual end-members of this soil-landscape continuum (Figure 5.2) are hillslopes where flow continuously converges (water-gathering) from summit to base versus hillslopes where flow continuously diverges (water-spreading) from summit to base.

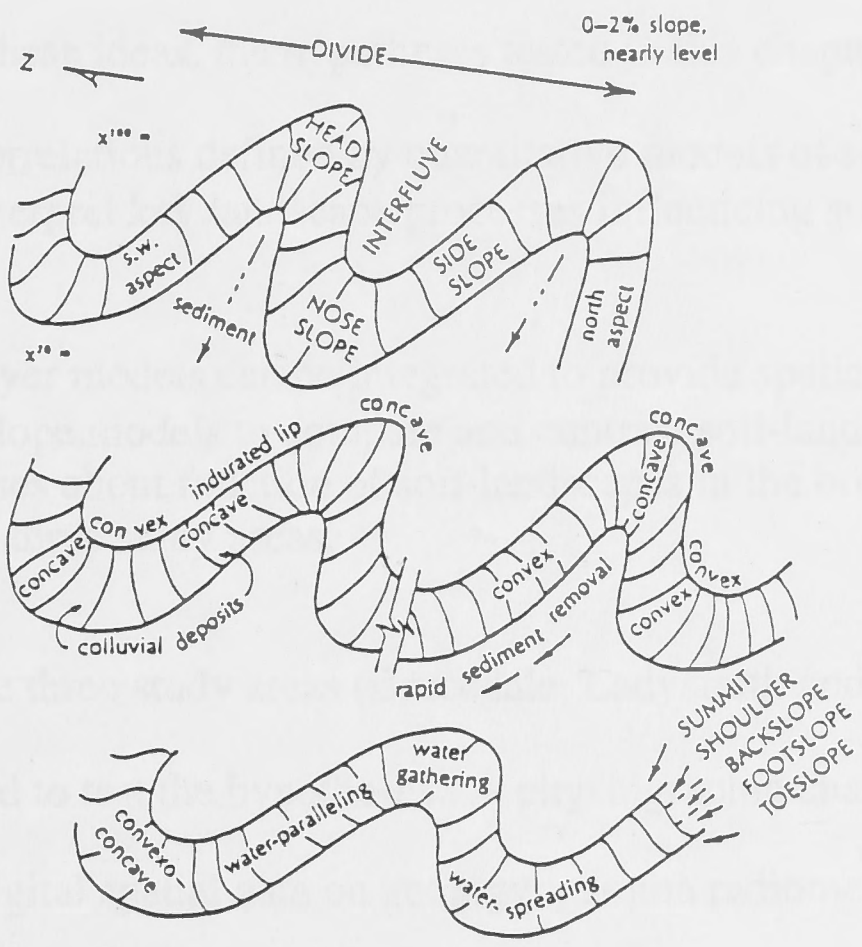


Figure 5.2 Catena Hillslope Model (from Buol *et al.*, 1989)

Some argue that development of soil taxonomic and mapping unit paradigms have caused a shift in emphasis away from an integrated understanding of soil-landscape continuum patterns and processes (Moore *et al.* 1993; McSweeney *et al.* 1995). Another obstacle has been the lack of quantitative tools to model catchment or watershed context, flow connectivity, adjacency and accumulation in three-dimensional landscapes. Digital terrain analysis methods have advanced considerably and Speight (1977), Moore *et al.* (1991; 1993), McSweeney *et al.* (1994) and others have presented tables indicating the process significance of various terrain attributes. Terrain analysis and other new quantitative techniques (e.g. statistical software, GIS) provide a broad array of tools that may be integrated to re-conceptualize how we conduct analyses for modelling and visualization of soil-landscapes in the broader environmental context. Quantitative modelling of soil layer patterns in three-dimensional landscapes incorporating watershed context, flow connectivity, adjacency and accumulation may provide a better framework for integrated soil-landscape analysis.

5.1.2 Hypotheses and Concepts

Building on these ideas, the hypotheses tested in this chapter are:

- environmental correlations defined by quantitative models of soil layer patterns can be used to interpret key landscape processes influencing soil layer development; and
- developed soil layer models can be integrated to provide spatially averaged and quantitative hillslope models to compare and contrast soil-landscape structure and develop hypotheses about function of soil-landscapes in the broader environmental context of the three study areas.

Data from the three study areas (Brucedale, Ladysmith and Griggward) shown in Figure 3.1 are used to test the hypotheses. A physiographic characterization (i_n) using all available digital spatial data on geology, gamma radiometrics, contemporary climate and topography is first provided to set the environmental context for soil layer and integrated hillslope modelling. The use of GIS and statistical modelling tools are

implicit throughout the chapter and enable the visualization of collected data and integrated hillslope models.

5.2 MATERIAL & METHODS

5.2.1 Study Area: Physiographic Characterizations

Graphics defining the relative environmental attribute spaces and gradients encompassed by the geographical extent of each study area are presented in Figures 5.3-5.5. The climatic and parent material characterizations define broader scale environmental variations at the upper meso-scales (i.e. the scale above intended application).

A simple topographic characterization is also presented to indicate basic landform differences between the study areas. Details of the specific sampling strategy for collection of soil core samples at the local hillslope scale (i.e. the scale of intended application) in each study area are provided in Section 3.2. The Brucedale study area has mixed agricultural land uses of cereal cropping and pastoral grazing. Ladysmith and Griggward are dominated by pastoral grazing land use.

Contemporary Climatic Characterization

Figure 5.3 displays the climatic characterization for the three study areas with a subset of the climatic variables generated by the ANUCLIM modelling package (McMahon *et al.* 1995) using a 245m grid spacing DEM and climate stations for the Wagga region. This upper meso-scale grid spacing does not capture local hillslope variation. Annual mean values of precipitation, radiation and temperature are provided to indicate basic study area differences. Variables that provide a simple understanding of water balance and comparison of climate during periods that may limit biological activity (e.g. moisture during hot summer, temperature during winter) are also provided. The data indicate that the three study areas straddle a climatic gradient where temperature decreases and moisture increases from north to south from Brucedale to Ladysmith and Griggward.

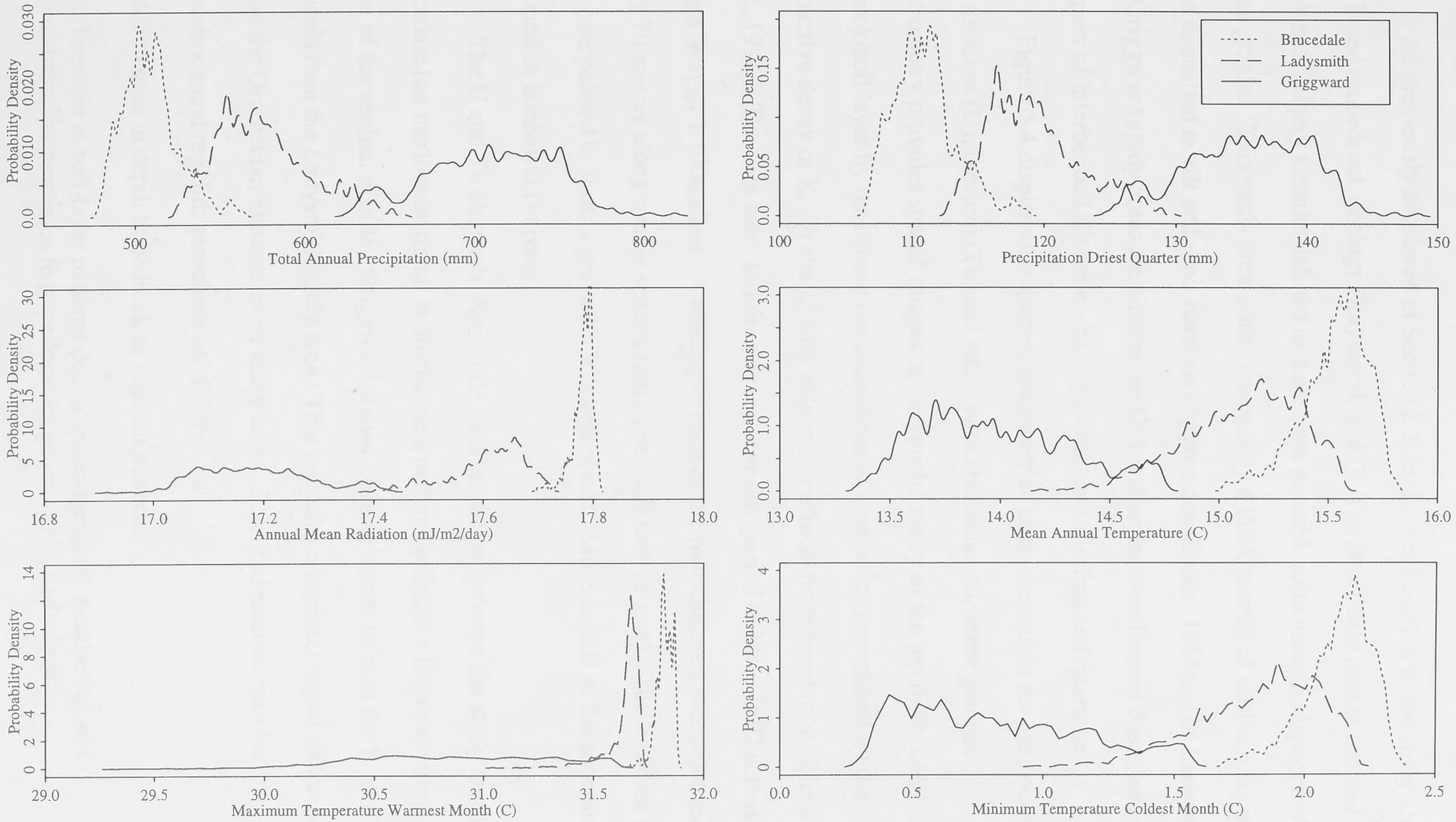


Figure 5.3 Contemporary Climatic Characterization

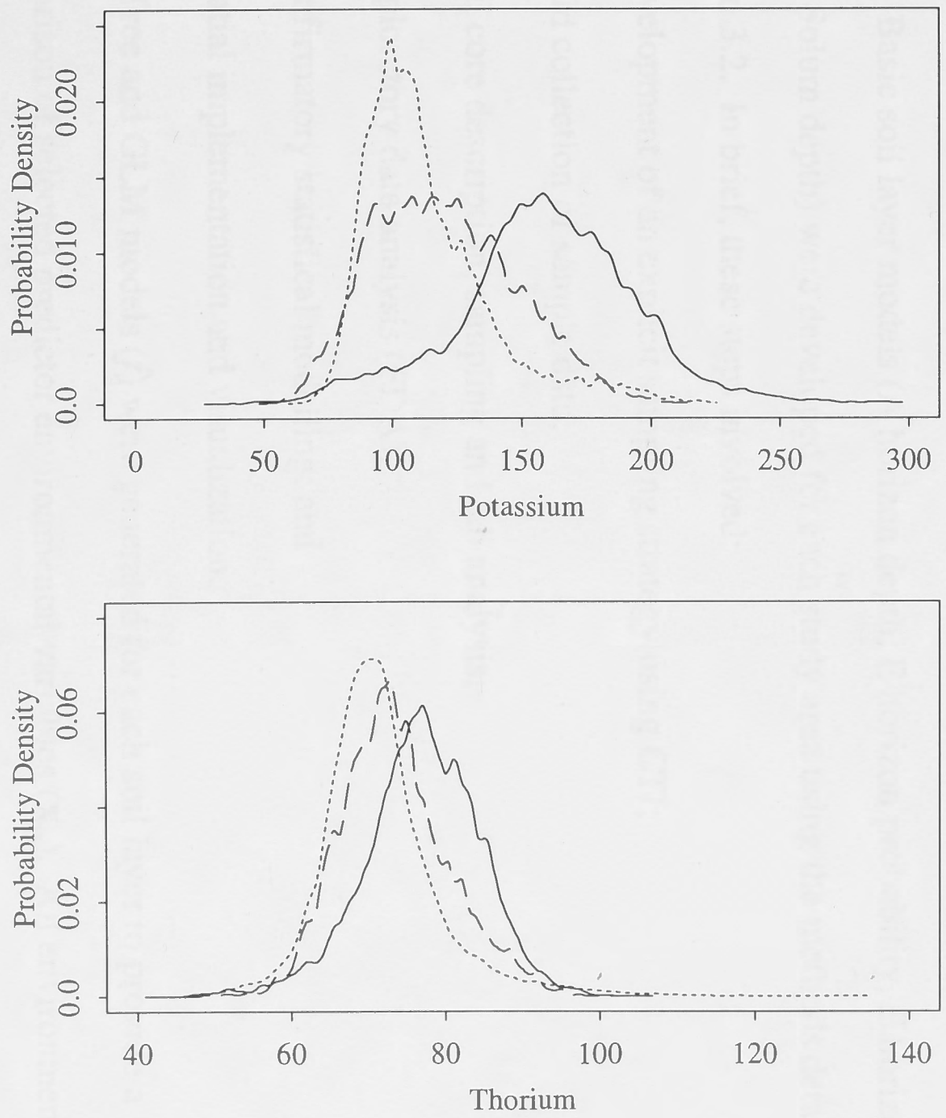
Parent Material Characterization

As previously discussed in Section 2.2, initial stratification of the study areas was based on bedrock geology (Raymond, 1992). The Brucedale area is situated on the Wantabadgery Granite defined as a medium grained, sometimes feldspar phyric, biotite granite. This study area is also covered by a thick mantle of aeolian clay or parna deposited as silt and sand sized clay aggregates (Butler, 1956). The Ladysmith and Griggward study areas are situated on Ordovician meta-sediments described as a complex of interbedded siltstone, shale, phyllite, minor schist and quartzite.

Figure 5.4 displays the gamma radiometrics characterization for the potassium (K), uranium (U), thorium (Th) and total count channels. Airborne gamma spectrometry provides spatial images of the geochemistry of the top 0.30-0.45m of the rock/soil layer by measuring the abundance of gamma rays produced by the radioactive decay of K, Th and U with only minor effects from vegetation (Bierwirth *et al.*, 1996). The data were acquired over a nine day period in May, 1992 across the Wagga Wagga 1:100k sheet. A reading in each of the four channels was collected every 70 metres along evenly spaced flight lines 400 metres apart. The line data were then interpolated to form a grid with 50 metre pixels. More details of the process can be found in Bierwirth (in prep.).

The U and Th channels show little discrimination between the study areas. Brucedale has the lowest signal in all channels perhaps reflecting the weathering status of the aeolian mantle. Griggward shows a higher signal in both the K and Th channels than the Ladysmith study area. This indicates a regional signal difference within the Ordovician metasediment study areas. Potential reasons may be:

- bedrock and regolith geo-chemical differences;
- differences in depth to bedrock or regolith material;
- differences in soil-layer patterns due to spatio-temporal weathering and redistribution processes (e.g. pedogenic, geomorphic, hydrologic)



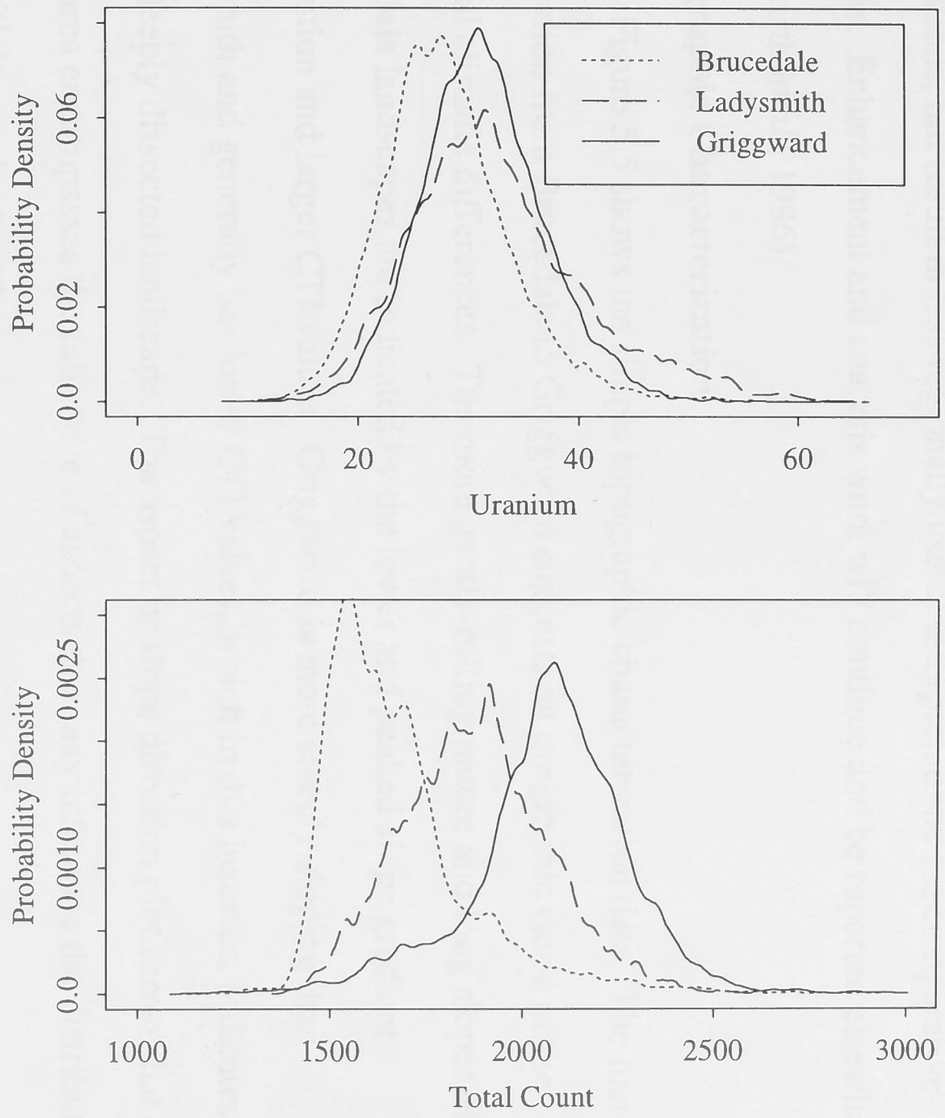


Figure 5.4 Radiometric Characterization

Radiometric image enhancement in each study area may improve soil pattern correlations, but result in less regionally useful interpretations for comparative analysis. Enhancement and analysis work will continue and be reported elsewhere (Bierwirth *et al.*, 1996)

Topographic Characterization

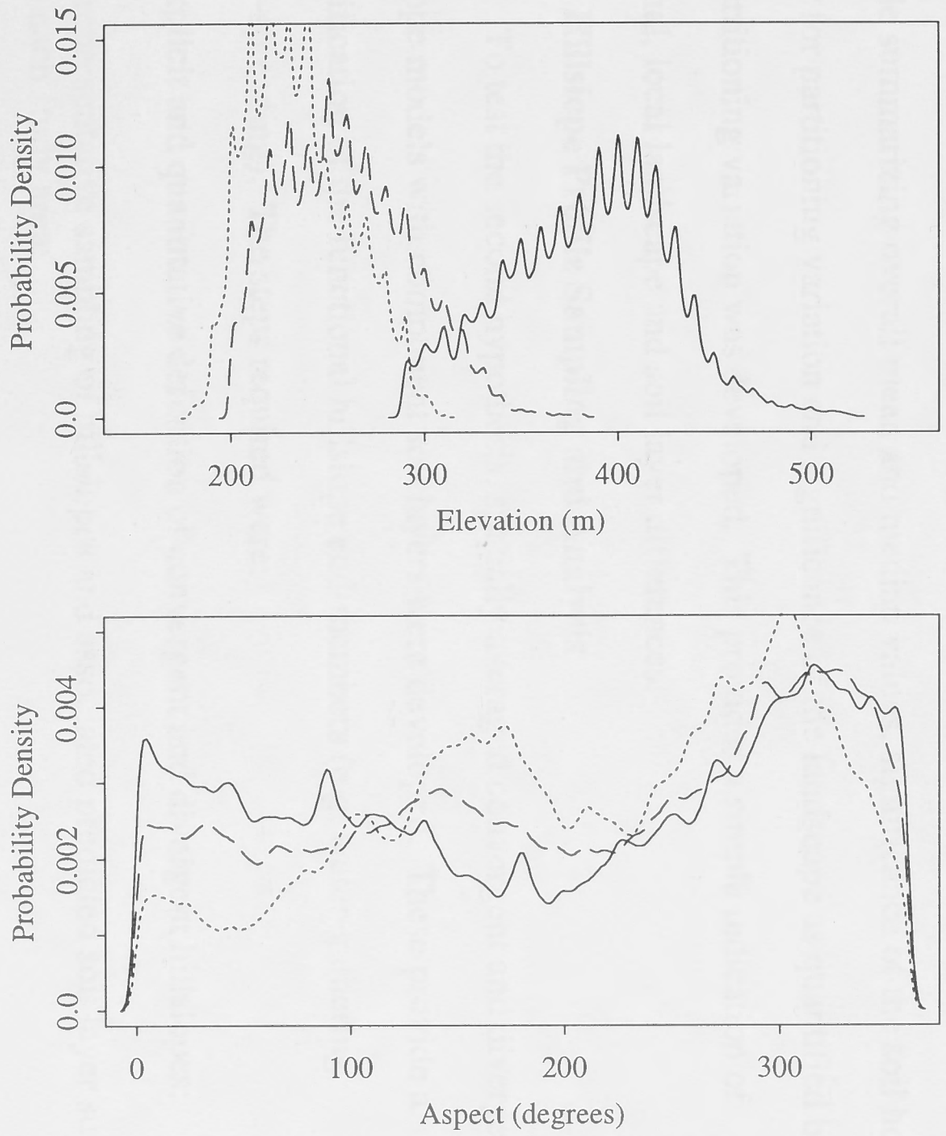
Figure 5.5 shows the simple topographic characterization data. The increase in elevation from Brucedale to Griggward suggests an orographic cause of the regional climatic differences. The more gently-rolling nature and long slopes of the Brucedale landscapes are indicated by the lower and peaked slope gradient distribution and larger CTI values. Griggward is more steeply sloping than Ladysmith and generally has lower CTI values, which in this instance, indicates a more deeply dissected landscape. The aspect or slope direction plot shows that each study area encompasses the full range of aspects that may influence the distribution of local hillslope solar radiation.

5.2.2 Statistical Modelling of Soil Layer Patterns

Basic soil layer models (A horizon depth, E horizon probability, E horizon depth, Solum depth) were developed for each study area using the methods detailed in Section 3.2. In brief, these steps involved:

- development of an explicit sampling strategy using CTI;
- field collection of sample data;
- soil core description, sampling and lab analysis;
- exploratory data analysis (EDA);
- confirmatory statistical modelling; and
- spatial implementation and visualization.

Both Tree and GLM models (f) were generated for each soil layer to provide a comparison of selected predictor environmental variables ($\mathbf{x_n}$). All environmental variables (climatic, gamma radiometric, terrain, orthophoto tone) were included as



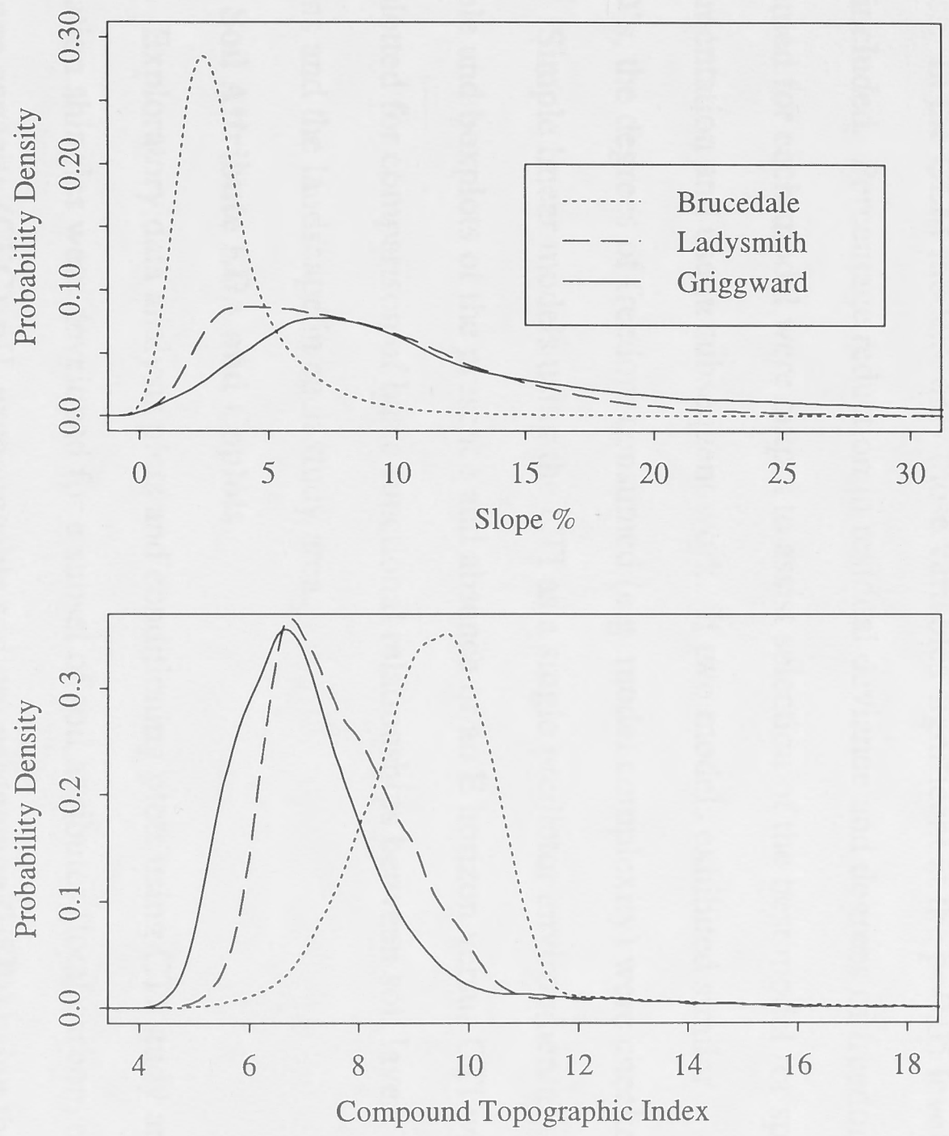


Figure 5.5 Topographic Characterization

potential predictors in the EDA and confirmatory statistical model development process. In the GLM models, only those variables significant at the p=0.05 level were included. Percentage reduction in residual deviance and degrees of freedom consumed for each model were output to assist selection of the best model for spatial implementation and use in subsequent work. If two models exhibited similar %RID's, the degrees of freedom consumed (e.g. model complexity) were checked.

Simple linear models using the CTI as a single predictor environmental variable and boxplots of the presence and absence of an E horizon versus CTI were also plotted for comparison of basic functional relationships between soil layer patterns and the landscape in each study area.

5.2.3 Soil Attribute EDA and Coplots

Exploratory data analysis plots and conditioning plots using CTI study area population shingles were developed for a subset of soil attributes (total carbon, cation exchange capacity (CEC), pH, exchangeable sodium percentage (ESP)) using the methods outlined in Section 3.2.6. The individual plots are presented in Appendix 1. A table summarizing overall mean and median values, significance of the soil horizon factor for partitioning variation and significance of the landscape as quantified by CTI for partitioning variation was developed. This provides a simple indication of regional, local landscape and soil layer differences.

5.2.4 Hillslope Profile Sampling and Analysis

To test the second hypothesis, spatially averaged convergent and divergent hillslope models with component soil layers were developed. These provide a quantification of the functional hillslope end-members (e.g. water-gathering, water-spreading). The steps required were:

- explicit and quantitative definition of convergent and divergent hillslopes;
- a representative sampling of hillslopes and associated predicted soil layer surfaces in each study area;
- collation, editing, analysis and standardization of the sampled predictive surfaces; and

- visualization of the quantitative hillslope models.

A hillslope was defined as a spatial object that maintains flow connectivity from summit (hillslope initiation) to base (hillslope conclusion). Following empirical experimentation with digital terrain attributes, hillslope initiation cells were defined as those cells with less than two DEM grid cells flowing into them. Hillslope conclusion cells were defined as cells with greater than 100 DEM grid cells flowing into them. These generally corresponded with ridge-lines and stream-lines as defined by the 1:25k topographic map sheets. By definition, convergent cells have plan curvature values greater than zero and divergent cells have plan curvature values less than zero. Flow connectivity was defined by placing flow vector arrows determined by a deterministic eight direction (d8) flow routing algorithm over the spatial display. A conditional map algebra 'if' statement was used to generate hillslope sampling grids meeting these criteria for each study area. This statement is as follows:

```
if (NCELL < 2)
        hill = 3  "hillslope initiation"
else if (NCELL > 100)
        hill = 2  "hillslope conclusion"
else if (PLCRV > 0)
        hill = 4  "convergent grid cell"
else if (PLCRV < 0)
        hill = 8  "divergent grid cell"
endif
```

Figure 5.6 shows a spatial display of these definitions for a small area at Griggward.

Sampling Procedure

Sampling of individual hillslope profiles was done by tracing and generating a line vector from a hillslope initiation node down the hillslope following connected flow vectors to a hillslope conclusion node. A convergent sample vector was derived by tracing connectivity through convergent nodes and vice versa for divergent nodes. The line vectors were then used to sample values along the vector from an elevation

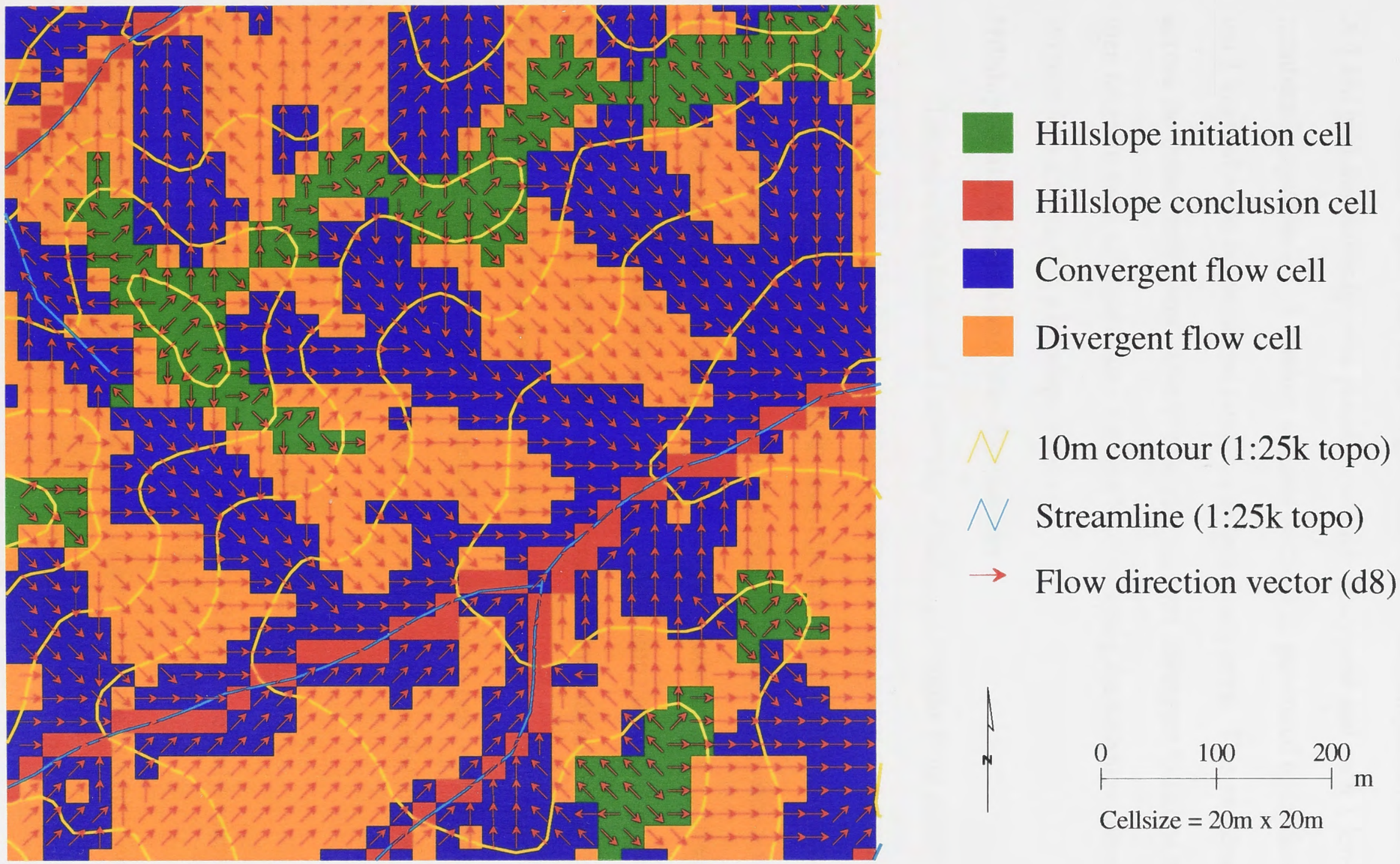


Figure 5.6 Hillslope Profile Sampling Model

(DEM) surface and predicted surfaces of A horizon depth, E horizon probability, E horizon depth and solum depth.

Ten convergent and ten divergent hillslopes were sampled in each study area. A 1 km grid line coverage was placed over each study area and each 1 km^2 cell numbered sequentially. A random number vector was generated to randomly select ten 1 km^2 subareas for hillslope sampling in each study area. This avoided bias across the three study areas. One convergent and one divergent hillslope sample was then taken in each selected 1 km^2 subarea by displaying the hillslope sampling grid (Figure 5.6) and tracing a hillslope vector.

Hillslope Data Analysis and Standardization

The sampling indicated a diversity of hillslope lengths from summit to base for each study area. The hillslope concept defined above is that of a spatial object that maintains both spatial connectivity and adjacency along the flow path. Adjacency meaning that the hillslope is an monotonic vector (e.g. top -> middle -> bottom). Therefore, an averaging of the ten profile values for a specific variable (e.g. A horizon depth) at equal hillslope distance intervals along the hillslope length will often be inappropriate because values at the same hillslope distance may be at different relative hillslope positions (i.e. the middle of a hillslope versus the bottom). Consequently, a standardized hillslope concept was developed to maintain spatial connectivity and adjacency for derivation of spatially averaged hillslopes. This involved the following steps:

- fit a spline model to the hillslope profile data for each sample;
- generate a predicted value from the spline model for 100 evenly-spaced increments along the entire length of the hillslope vector;
- take the mean of the ten predicted values of each hillslope profile at each equivalent increment (1...100) to generate a mean hillslope vector for each predicted surface (elevation, A horizon depth, E horizon probability, E horizon depth, solum depth); and
- use the average hillslope length from the ten samples as the standardized hillslope length for visualization of the spatially averaged hillslope.

A function was written using the Splus language (Statistical Sciences, 1993) to automate the steps outlined and output a dataframe containing the mean values for the component soil layer surfaces for each spatially averaged hillslope.

Visualization

Integrated hillslope visualizations were developed by successively plotting cross-sections of solum depth, E horizon depth and A horizon depth using the hillslope elevation as the surface . Cross-section fill colours were selected to closely match the true soil colours as determined from the sample data. E horizon depth values were conditionally plotted only for those hillslope distances where the probability of an E horizon was 0.5 or greater. The graphical axes (x = hillslope distance, y = hillslope height) were established to fit all plots on any particular visualization to highlight relative differences. The soil layer depths were multiplied by a factor of ten to enhance cross-section soil layer display, and the ordinate extended to negative hillslope heights to fit the soil layers on to the display. Hence, an increment of ten metres on the ordinate is equivalent to one metre of soil depth. The same process was also used for visualizing individual hillslope soil layer samples.

5.3 RESULTS & DISCUSSION

5.3.1 Soil Layer Models

Table 5.1 summarizes the soil layer models and explanatory environmental variables listed in relative order of inclusion in the model (i.e. largest %RID first). Table 3.1 (Section 3.2) provides a key to the abbreviated explanatory environmental variables selected and used in the models of Table 5.1. Percentage reduction in residual deviance ranged from 15 (Ladysmith E horizon Depth GLM) to 94 (E horizon probability Tree - Griggward, Ladysmith). No outliers or high leverage points were removed. The models indicate a broad diversity of predictability in the study areas,

Table 5.1 Study Area Soil Layer Models

response variable **s**	~	model type *f*	explanatory environmental variables ($x_1, \ldots x_n$)	%RID (d.f. consumed)
A Horizon Depth				
Brucedale		GLM	SLPP, TH400, PLCRV, MPLCRV, NCELL	31 (6)
		Tree*	FPL, CTI, MTCRV, K400, MPRCRV, NCELL, ELEV, SLPP, AMR	60 (9)*
Ladysmith		GLM*	CTI	61 (2)*
		Tree	CTI, K400, SPI, STRIN, ASP, SLPP, FPL	78 (7)
Griggward		GLM*	log(NCELL)	42 (2)*
		Tree	STRIN, NCELL, MPRCRV, MPLCRV, U400, PRCRV	78 (6)
E Horizon Probability				
Ladysmith		$GLM_{(logistic)}$	SLPP, MSLP, MTCRV	40 (4)
		Tree*	CTI, NCELL, TH400, MPRCRV, AMR, MPLCRV, SLPP	94 (7)*
Griggward		$GLM_{(logistic)}$	FPL, MPLCRV, MTCRV	36 (4)
		Tree*	MTCRV, PDQ, FPL, MPLCRV, K400, PRCRV	94 (6)*
E Horizon Depth				
Ladysmith		GLM	TCRV, ASP	15 (3)
		TREE*	CTI, TCRV, ASP, PDQ	59 (4)*
Griggward		GLM*	PLCRV, RED, GREEN	63 (4)*
		Tree	SPI, AMR, CTI	61 (3)
Solum Depth				
Brucedale		GLM	CTI, NCELL, TCRV, PLCRV	36 (5)
		TREE*	CTI, PLCRV, SLPP, FPL, NCELL, PRCRV, SPI	81 (7)*
Ladysmith		GLM	CTI, K400, SPI	75 (4)
		Tree*	CTI, MPRCRV, K400, TCRV, ELEV	90 (5)*
Griggward		GLM*	log(NCELL)	42 (2)*
		Tree	CTI, NCELL, K400, TCRV, STRIN, PRCRV, AMR	77 (7)

* denotes model chosen for spatial implementation

with the Tree models consistently providing better predictions (based on %RID), but consuming more degrees of freedom by including more explanatory variables.

Environmental Variables: Relative Usefulness

In every model, a terrain variable provided the largest reduction in deviance and a broad range of terrain attributes were shown to be useful. This suggests that lower meso-scale (20m) terrain attributes are capturing variation relating to hillslope processes controlling soil layer patterns. CTI and it's components (slope and flow accumulation (NCELL)) were, overall, the most useful terrain attributes. It may be suggested this is because of the use of CTI as a provisional model for sample allocation. However, plotting of sample allocations in other environmental attribute spaces (terrain, climate, radiometric) indicated that other environmental attribute ranges and distributions were generally well covered by the sampling (n=~70). This may not be the case with smaller sample numbers (n<30). In many of the Tree models, CTI was incorporated at more than one node.

Contour curvature and the upslope area contour curvature variables (PLCRV, TCRV, MTCRV, MPLCRV) were also useful, indicating that local and contextual flow convergence and divergence relates strongly to soil layer patterns.

Potassium and thorium gamma radiometric variables were useful, usually as a second or third variable incorporated after a terrain attribute. This suggests a complementary role for radiometrics in capturing mesoscale geochemical variations important for modelling soil layer patterns. Uranium was selected only once at a low Tree node. Those explanatory variables included at the low Tree nodes, or as the later variables in the model, often provide little reduction in residual deviance.

Annual mean radiation was the best climatic variable (aspect was also incorporated three times) suggesting that solar radiation, or on a process basis, water balance is important in these landscapes and that computation of more local hillslope radiation variables (e.g. 20m grid spacing) may be beneficial for soil layer prediction and understanding of the energy balance.

Process Interpretation

A-horizon depth is a good indicator of biological productivity, stability with respect to surface erosional processes, or mass balance of erosional and depositional processes. For A horizon depth, the predictors in Griggward and Ladysmith are almost exclusively contextual and secondary terrain attributes. Brucedale predictors were dominated more by local variables (e.g. slope gradient) suggesting that hillslope processes are not as important.

E horizon probability and depth are commonly interpreted as indicators of lateral throughflow, leaching, or waterlogging, particularly in texture contrast soils. The aeolian clay derived soils in the Brucedale area exhibit only gradual variation in texture from the A to the B horizon. The B horizons are also well-structured. This suggests there are no impediments to the vertical movement of water through the soils in these landscapes. The sampling showed only two occurrences of E horizons, therefore E horizon models were not developed for Brucedale. Predictors of Griggward E horizon probability and depth indicate that water flow processes and flow convergence and divergence are critical. Slope and tan curvature were important in Ladysmith perhaps indicating that local energy is more important than flow accumulation. E horizon depth models in both Ladysmith and Griggward incorporated climatic (AMR, PDQ), slope azimuth (ASP) and red and green orthophoto bands. Individual scatterplots show relationships where wetter sites supporting healthy vegetation and/or high levels of organic matter have deeper E horizons. This suggests a biophysical and related pedogenic process of podzolization (e.g. losses) more prominent in favorable hillslope positions.

Solum depth is an overall indicator of stability to hillslope erosional and depositional processes and *in-situ* soil forming processes. CTI, flow accumulation and contextual terrain attributes were the best predictors in each study area for solum depth suggesting that hillslope process connectivity is important and well captured by these meso-scale terrain variables.

Overall, it is difficult to comprehensively interpret landscape processes from the Tree and GLM models because different environmental variables were often selected for the same study area. The first variable in the Tree models were often a secondary terrain or contextual attribute, whereas the GLM models tended more towards local primary terrain variables. This highlights an influence of the modelling and attribute selection technique chosen. The Tree modelling approach iteratively splits the data in search of the most homogeneous (e.g. lowest variance) subsets, whereas GLM methods attempt to reduce the overall variance with a fitted term. Thus the GLM methods may miss important conditional relationships in the data suggestive of different generative processes (e.g. erosional process zones versus depositional).

A simpler way to visualize comparative relationships is to fit single predictor GLM's to illustrate relationships between the soil layer attributes and explanatory environmental variables. Figure 5.7 shows CTI models of A horizon depth and solum depth with individual sample points along with boxplots of CTI versus E horizon presence/absence for Ladysmith and Griggward. The %RID of each GLM model is also reported. The A horizon depth models and data indicate little variation in depth with landscape in Brucedale. Although the model slope is approximately the same for Ladysmith and Griggward, the intercepts and offset indicate that A horizons are deeper throughout the landscape at Griggward. However, the explained sample variation is significantly higher (%RID= 62 vs. 45) for Ladysmith suggesting a more complex suite of processes influencing A depth in Griggward.

The fitted model lines of solum depth versus CTI are remarkably similar, although the Brucedale depths are clustered in both CTI and solum depth space reflected by the low RID (24). The fitted solum depth model line at the low CTI's (high landscape) are almost identical for Ladysmith and Griggward, but the Ladysmith soils increase in depth more rapidly with CTI towards the lower landscape positions. The model fit is, again, tighter for the Ladysmith data. Variation about the fit, as indicated by the sample points, appears to increase after CTI of approximately 7.0

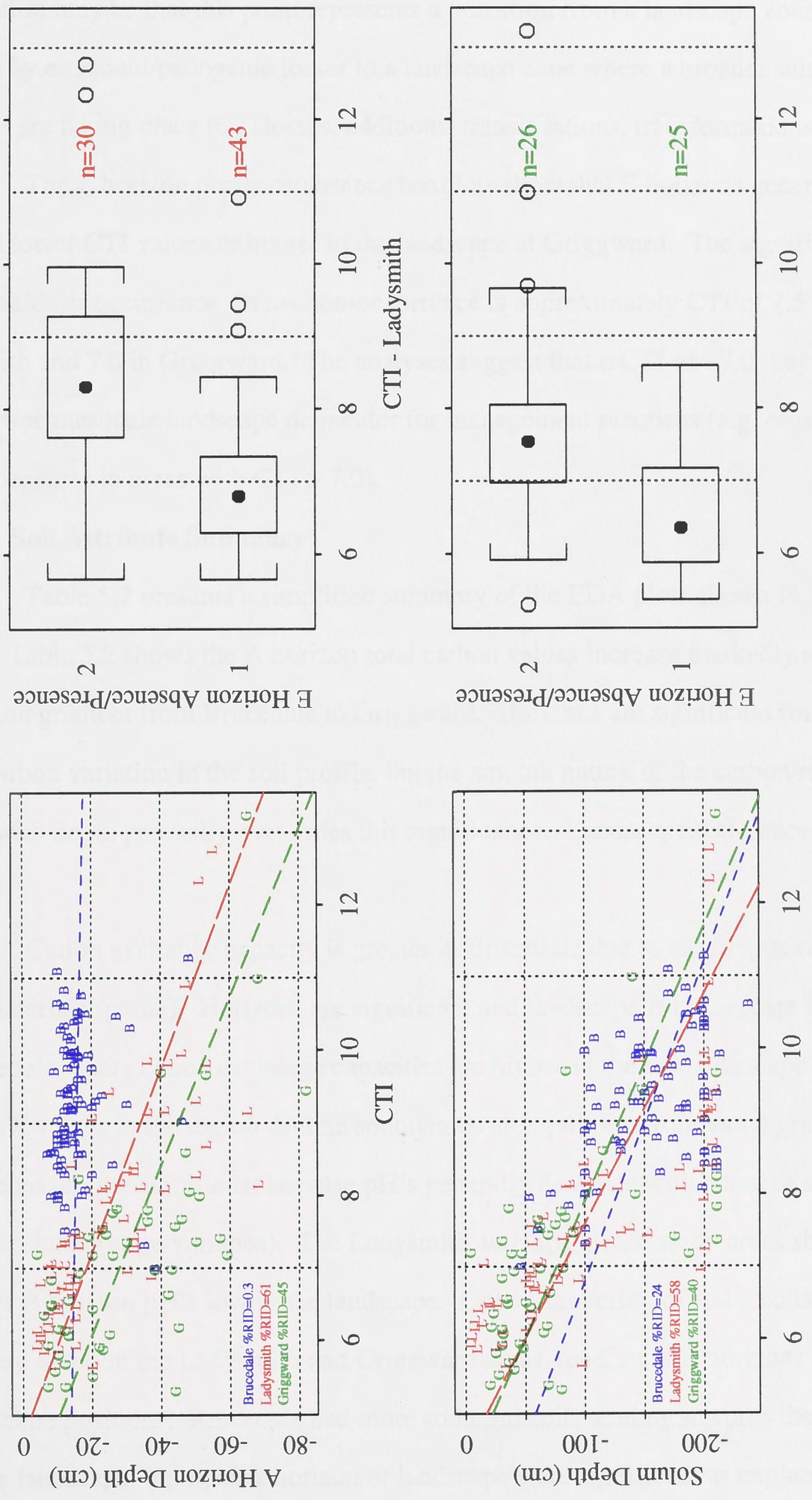


Figure 5.7 Soil Layer Versus CTI Comparisons

for both A horizon and solum depths in Ladysmith and Griggward. A possible explanation may be that this point represents a transition from a landscape zone dominated by erosional/pedogenic losses to a landscape zone where a broader suite of processes are taking place (e.g. losses, additions, translocations, transformations).

The E horizon presence/absence boxplots show that E horizons generally occur at lower CTI values or higher in the landscape at Griggward. The significant threshold for occurrence versus non-occurrence is approximately CTI of 7.5 in Ladymsith and 7.0 in Griggward. The analyses suggest that a CTI of ~7.0 may be a useful lower-mesocale landscape delineator for management practices (e.g. erosion control measures in areas with CTI > 7.0).

5.3.2 Soil Attribute Summary

Table 5.2 presents a simplified summary of the EDA plots shown in Appendix One. Table 5.2 shows the A horizon total carbon values increase markedly along the climatic gradient from Brucedale to Griggward. Horizons are significant for partioning carbon variation in the soil profile, but the smooth nature of the carbon relationship with depth generally over-rides this significance. Landscape differences are negligible.

Cation exchange capacity is greater in Brucedale due to the finer textured parent materials (parna). Horizons are significant and landscape is not, except for Brucedale where cation exchange capacities are higher in the low landscape positions. The pH values in the region decline southwards along the environmental gradient. Horizons are not significant because pH's generally decrease with depth of sample in a broad band (large variance). The Ladysmith and Griggward study areas show higher B horizon pH's low in the landscape. Only a scattering of sodic soils (ESP > 5) were noted in the Ladysmith and Griggward areas, tending to B horizons in low landscape positions. Brucedale had more sodic subsoils tending towards the middle of the landscape, but neither horizon or landscape were significant as explanatory predictors.

Table 5.2 Soil Attribute Summary

	mean	median	horizon	landscape
Total Carbon (A horizon)				
Brucedale	1.3	1.275	yes	no
Ladysmith	2.27	1.8	yes	no
Griggward	2.58	2.49	yes	no
CEC				
Brucedale	11.8	10.3	yes	yes
Ladysmith	7.8	7.5	yes	no
Griggward	7.6	7.4	yes	no
pH				
Brucedale	6.62	6.56	no	no
Ladysmith	6.09	5.96	no	yes
Griggward	5.68	5.61	no	yes
ESP				
Brucedale	2.13	1.0	no	no
Ladysmith	1.76	0.9	no	no
Griggward	1.49	0.9	no	no

5.3.3 Integrated Mean Hillslope Models

Figures 5.8-5.10 display hillslope sample surfaces and the spatially averaged mean divergent and convergent hillslope soil layer models for each study area. The sample surfaces quantify the hillslope heights, distances, shapes and overall hillslope diversity represented by the sampling. The percentage of the study area occupied by hillslope summit, convergent, divergent and hillslope conclusion cells is also reported. The integrated hillslope soil layer models provide a quantitative visualization of the soil layer patterns representating the end-members of the three-dimensional soil-landscape continuum within each area. They represent differences due to landform morphology and hypothesized local hydrological and geomorphic processes that laterally re-distribute water and soil material. Figures 5.11 and 5.12 provide a comparative visualization of the inter study area hillslope patterns more useful for interpreting regional differences due to broader environmental processes.

Intra Study Area Interpretations

Figure 5.8a shows that the Brucedale divergent hillslope samples occupy a wider envelope exhibiting greater diversity than the convergent samples and that the hillslope lengths are long, averaging around 580 metres with a drop in height of about 21 metres. The A horizon depths and solum depths are deeper on the convergent hillslope (Figure 5.8b), but there is not much difference with the divergent hillslope (Figure 5.8c). The consistent nature of the A horizon depth may be influenced by the widespread cereal cropping landuse in this study area where cultivation maintains a plow layer A horizon (Ap). Soil texture does not strongly contrast between the A horizon (clay loam) and B horizon (light clay). Field samples also indicated that the B horizons are well structured and permeable. These factors, coupled with the gently inclining topography and dryer and warmer climate, compared to Ladysmith and Griggward, suggest that surface and subsurface lateral flow are not as important in the Brucedale landscapes. This implies that the principal material and energy flow pathway in this landscape is in-situ vertical infiltration of water where it is either used by plants or lost to deep percolation.

Both hillslope models decrease markedly in solum depth close to the hillslope conclusion. Evaluation of the predicted Tree model surface suggests a limitation of the model based on an inappropriate branch and predictive surface step.

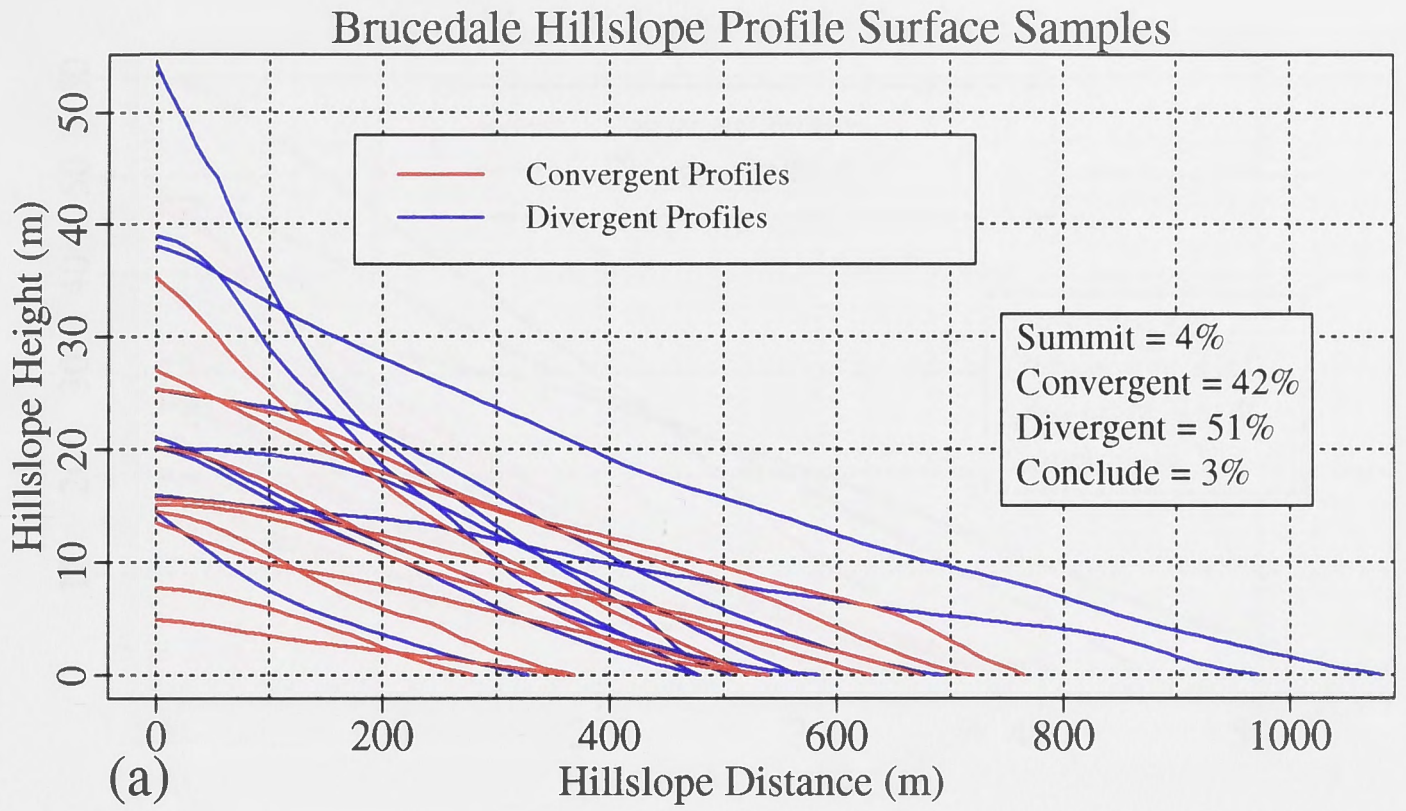
Figure 5.9a shows that the Ladysmith profile samples also show a broader envelope for the divergent samples and have an average hillslope length of 360 metres with average vertical drop of 30 metres. The soil textures of the A and B horizon contrast markedly in the Ordovician metasediment landscapes. The A horizons are typically loams and the B horizons, light clay to clay textures. The convergent and divergent hillslope differences (Figures 5.9a, b) are more pronounced than Brucedale. While the soil pattern is similar at the hillslope summit (Figures 5.9a, b), within 75 metres down hillslope, solum depths on convergent hillslopes begin to increase and E horizons occur. The overall convergent hillslope profile is more concave. One

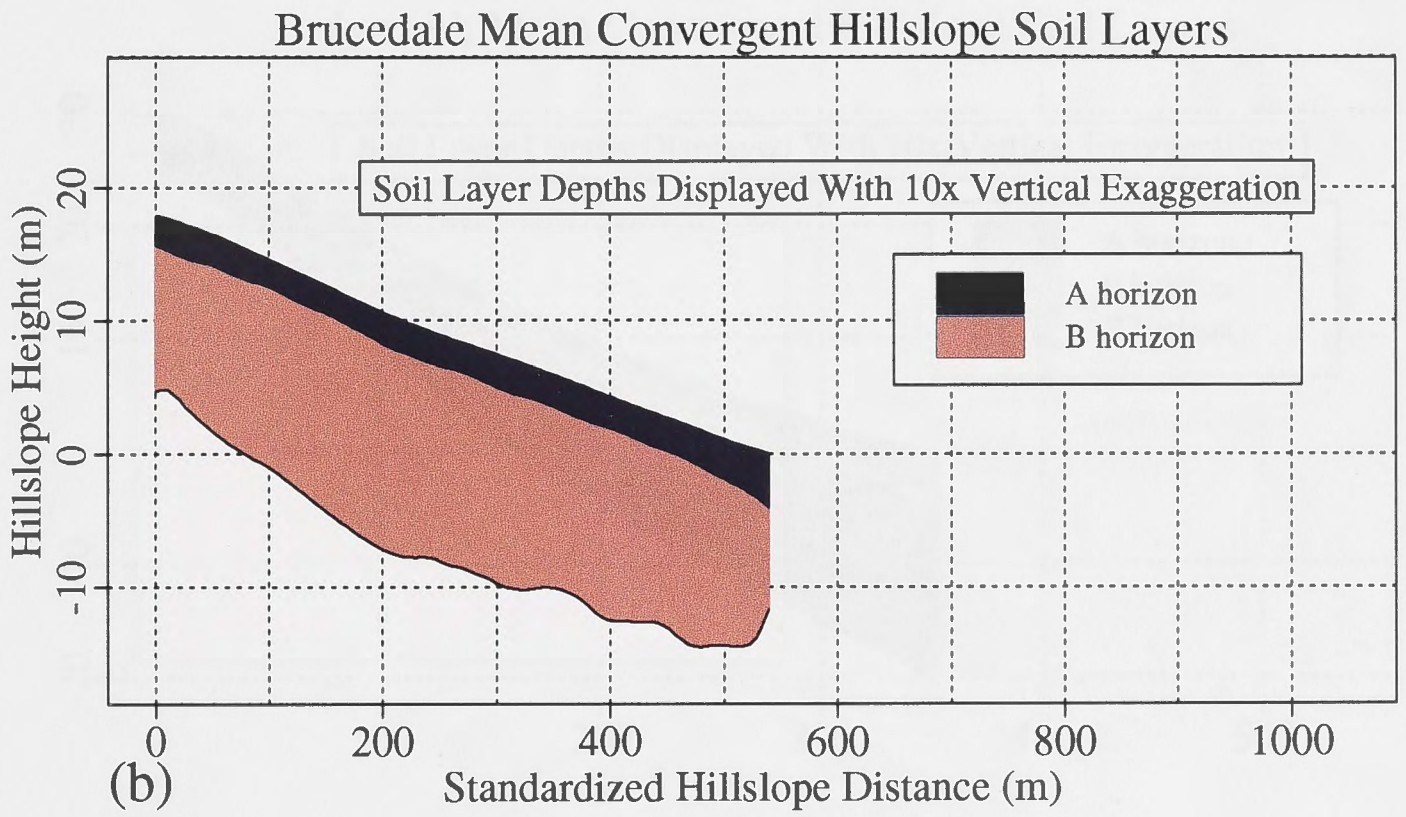
interpretation may be that the local re-distribution of surface and subsurface water is adding material below the summit by depositional processes and more frequently saturating the A horizon in this higher slope gradient environment causing subsurface lateral flow above the B horizon that removes solutes and colloidal material. As the hillslope approaches the conclusion point, E horizons become slightly shallower, while A horizon and solum depths are deepest. The connection of the hillslope to the broader catchment or watershed system where flow accumulation is much greater may mean that overland flow and alluvial processes are more dominant at the hillslope base.

The A horizon and solum depths are mostly constant along the divergent hillslope, with a marginally increasing solum depth, until a slight concave inflection point around 300 metres. At this point, solum depth increases and E horizons begin to occur.

The Griggward profile samples (Figure 5.10a-b) have an average hillslope length of 290 metres with an average vertical drop of 32 metres, indicating a higher energy environment than Ladysmith. The convergent and divergent relationship between overall solum depth and hillslope shape is similar to Ladysmith where depths are similar towards the top then digress at about 50-75 metres. However, very shallow E horizons begin to occur almost at the top of the divergent hillslope (Figure 5.10c) and gradually increase in depth down the hillslope. The occurrence of E horizons almost throughout the landscape in Griggward does not correspond with the boxplot of Figure 5.7 where the sample data show an approximate even split in probability of E occurrence. However, the E horizon probability models in Table 5.1 show that CTI was not used as a predictor and Figure 5.10 indicates a much larger proportion of the study area classified as hillslope initiation cells where E horizons are less likely according to the hillslope visualizations.

E horizons do not occur towards the crest of the convergent hillslope. The convex nature of the convergent hillslope summit may indicate a higher energy zone





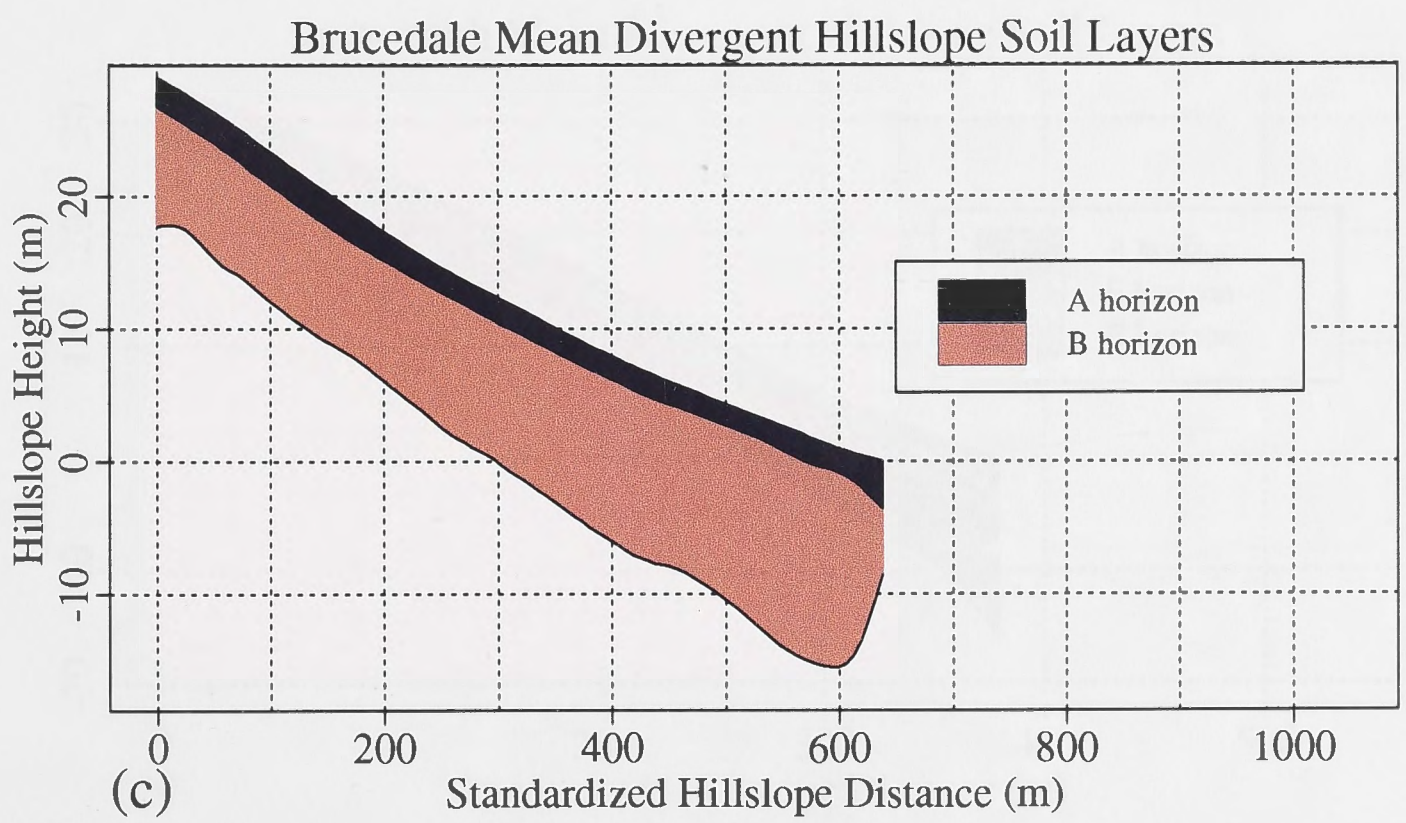
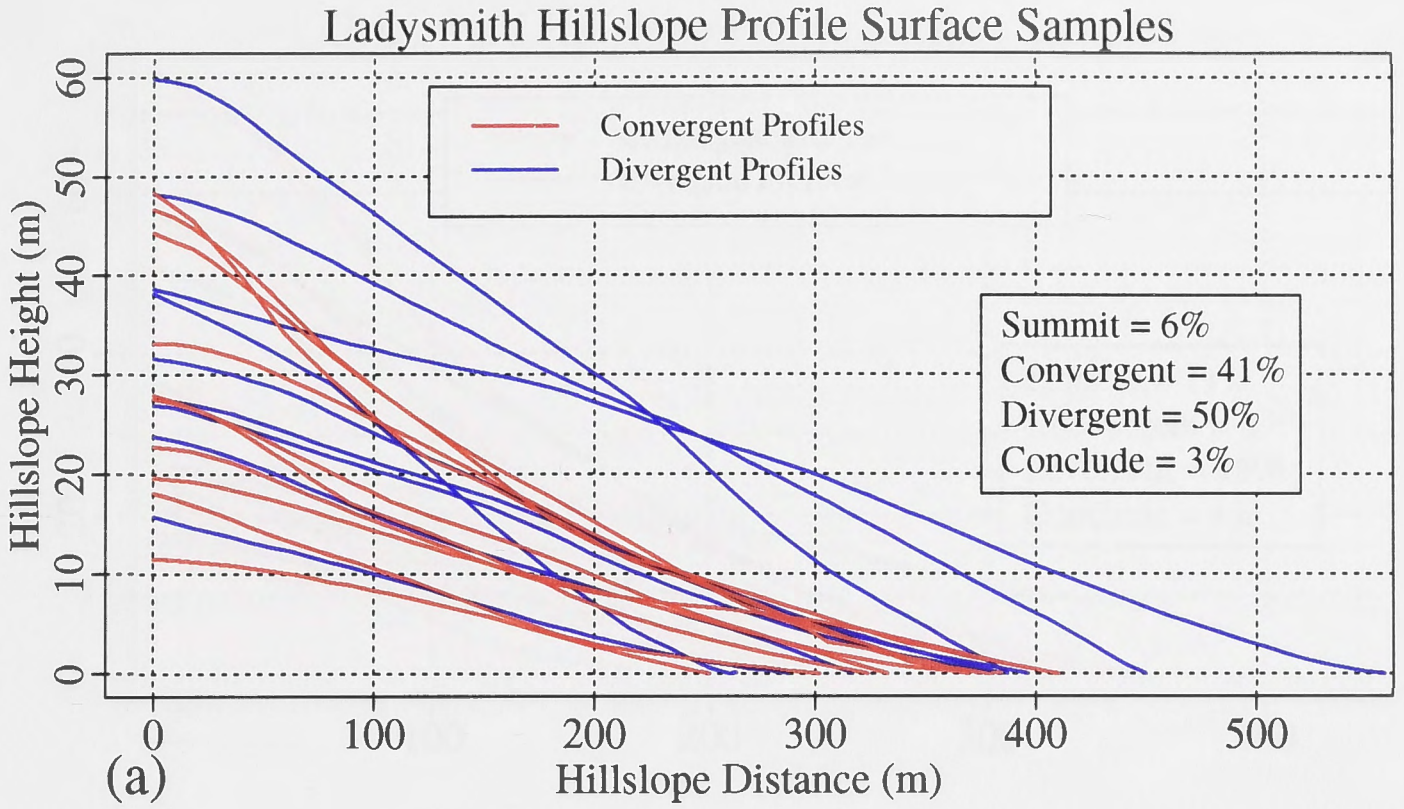
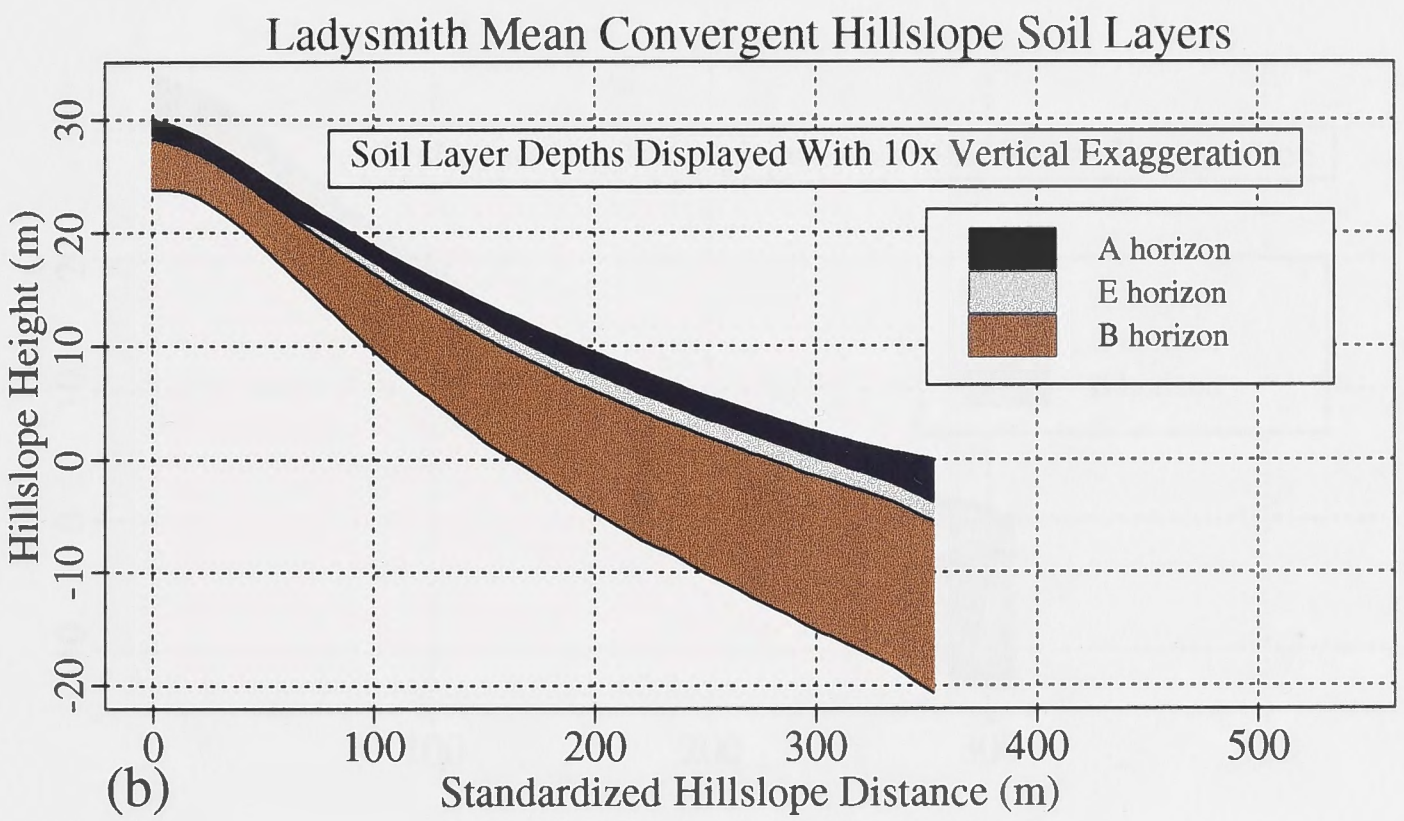


Figure 5.8 Brucedale Hillslope Models





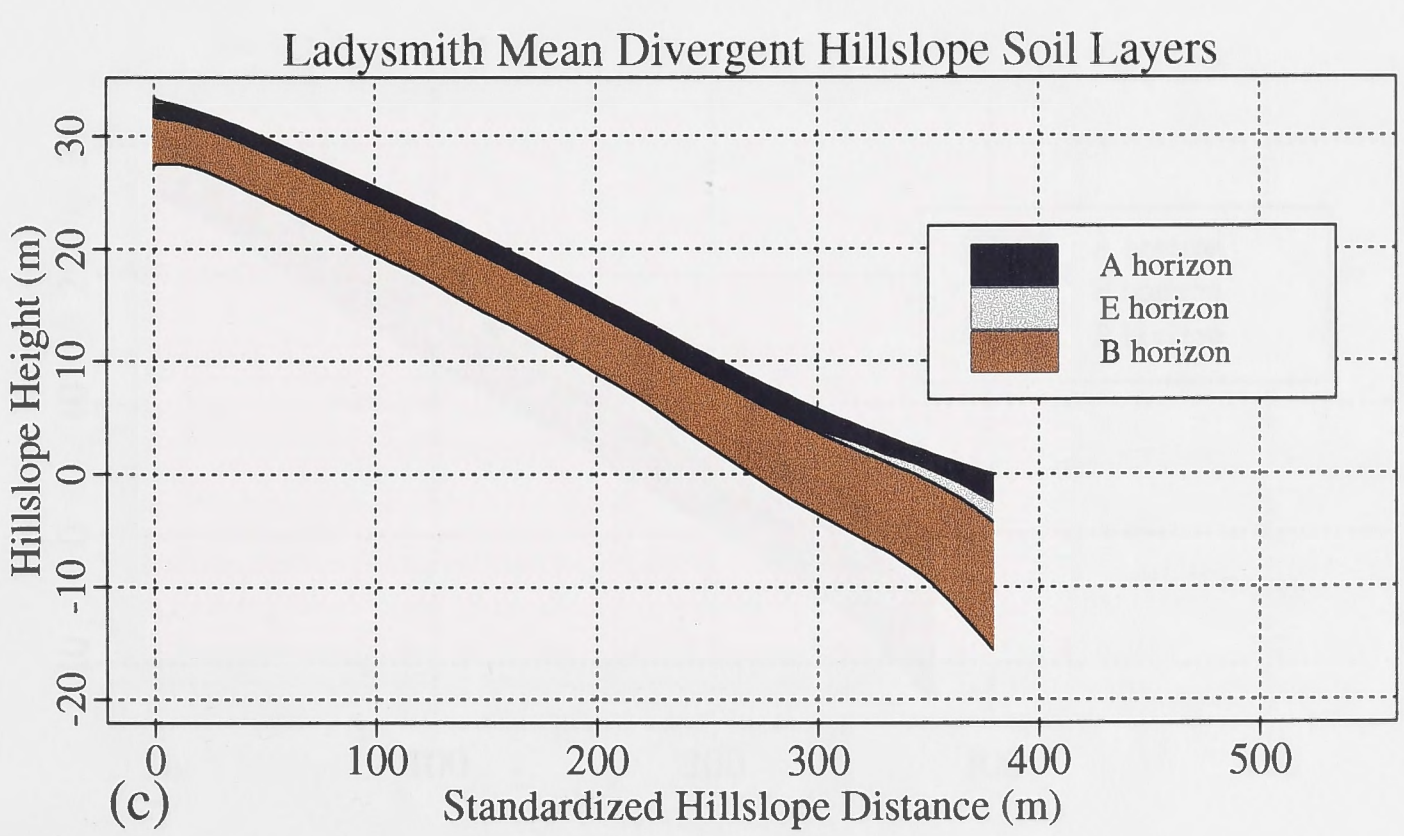
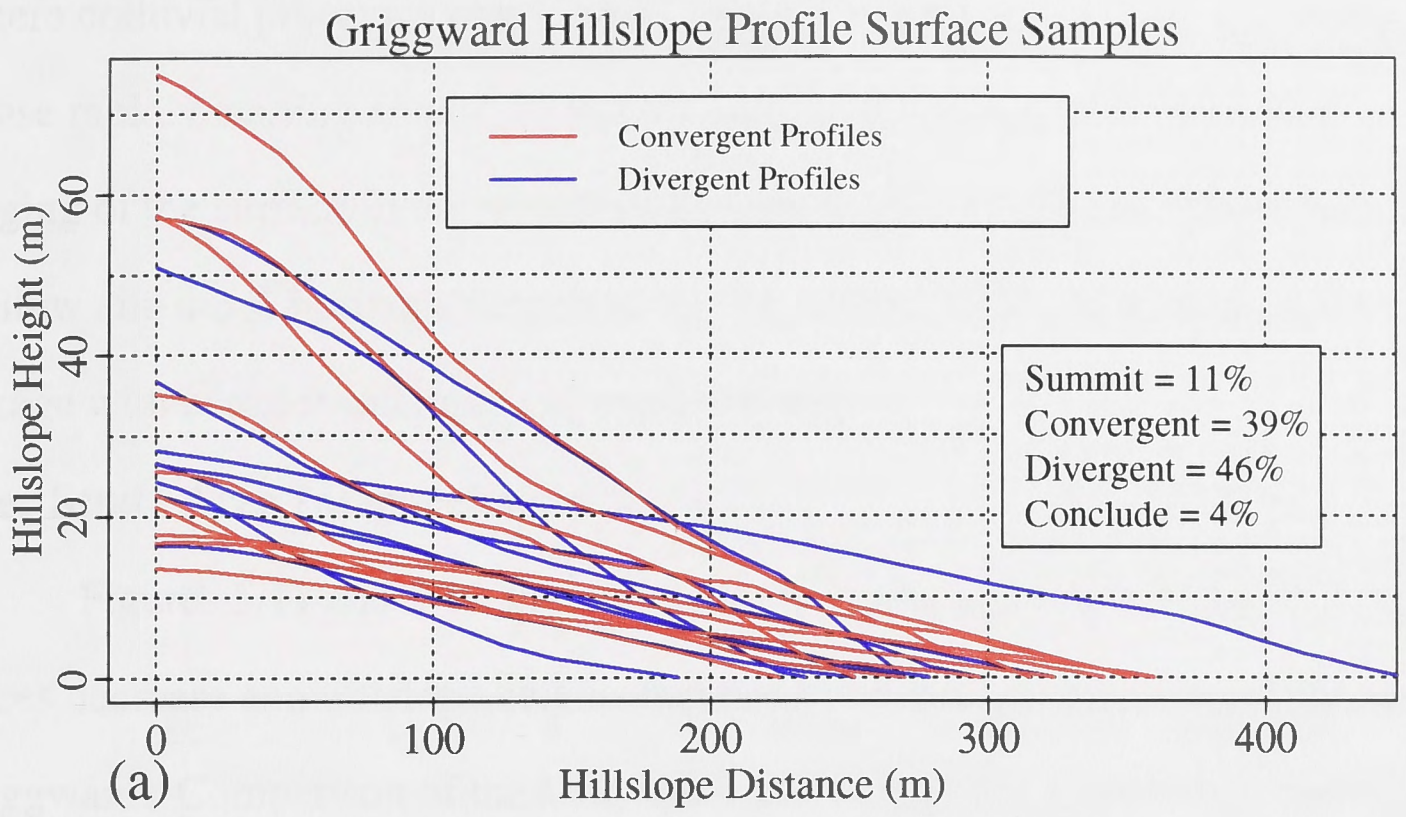
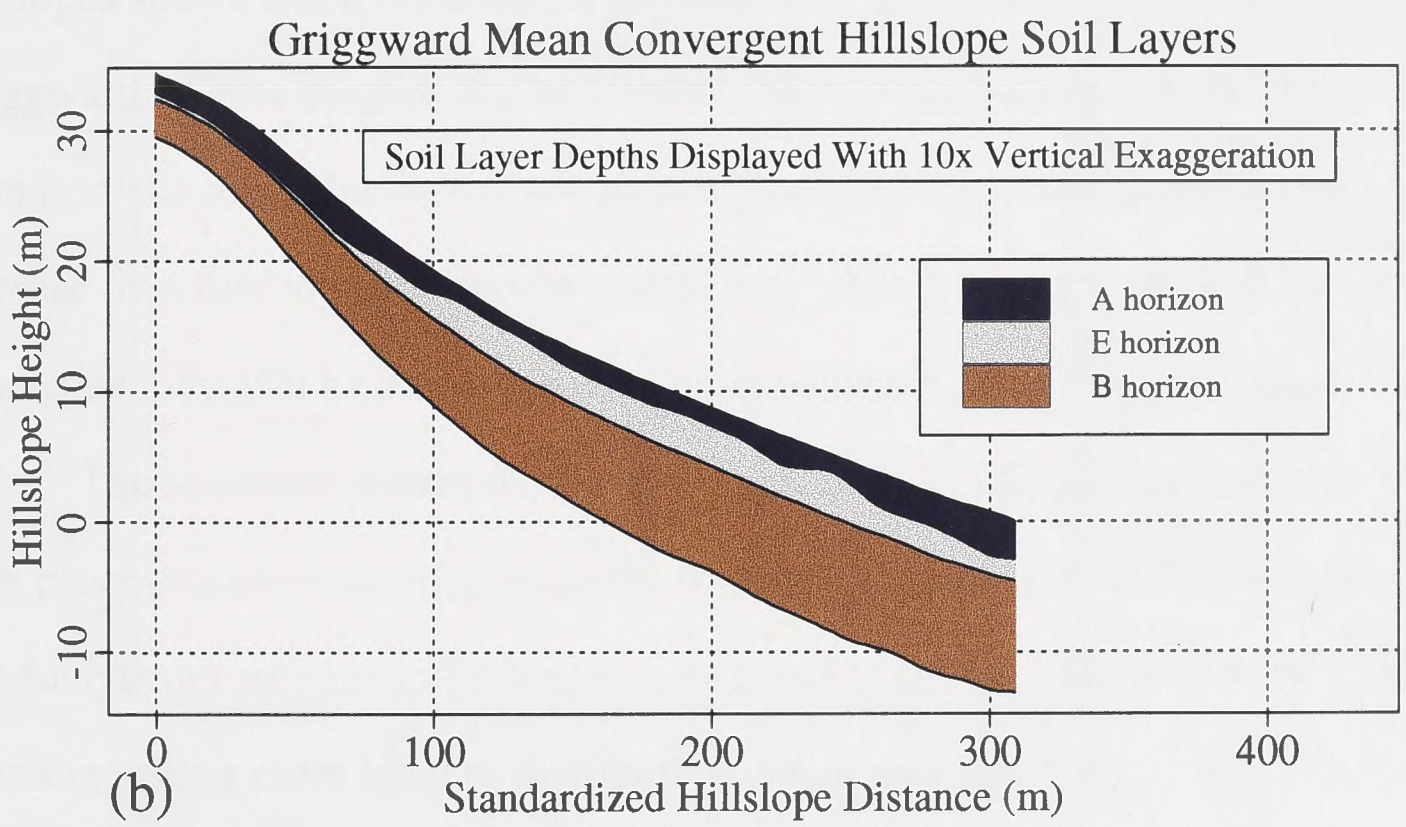


Figure 5.9 Ladysmith Hillslope Models





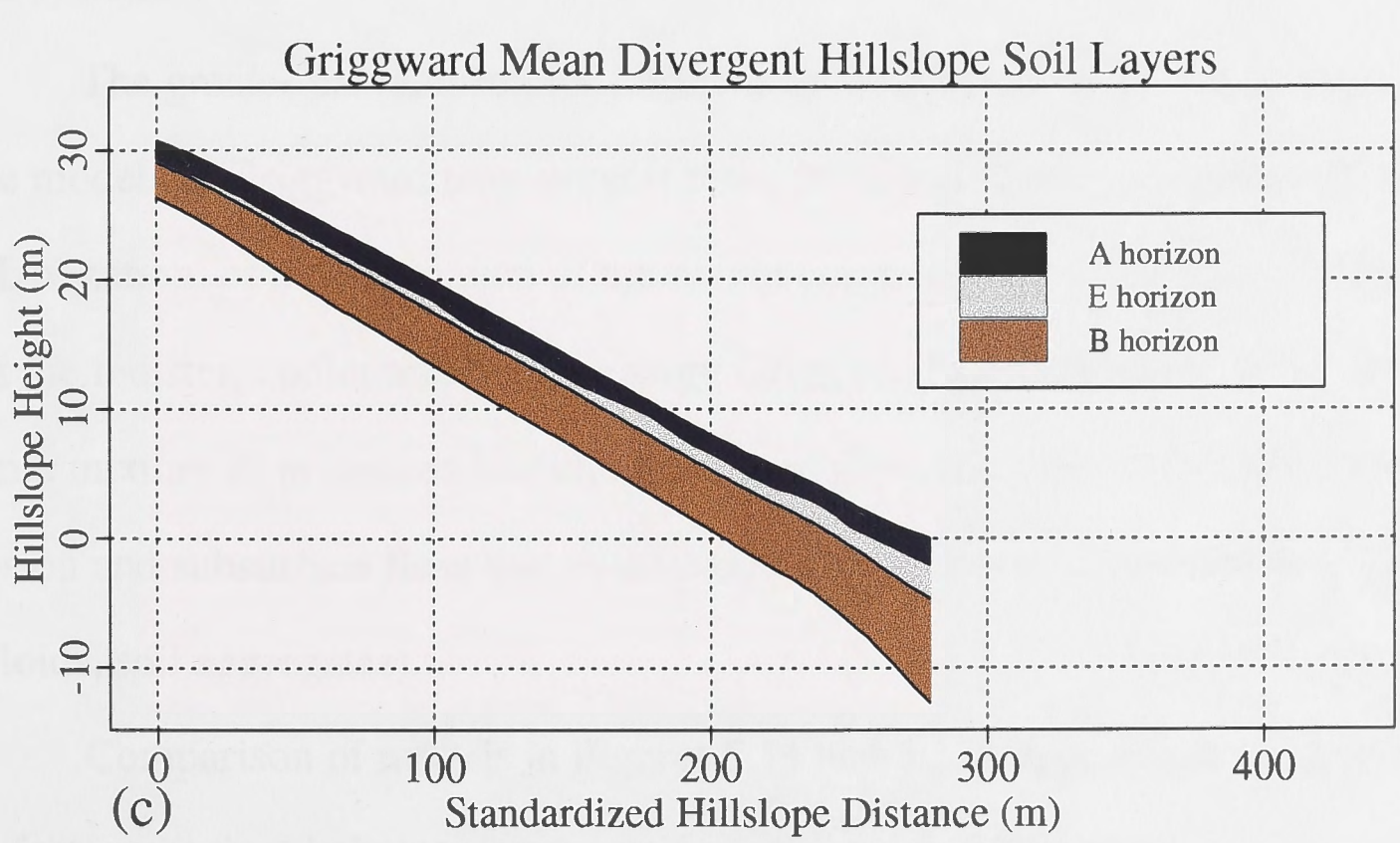


Figure 5.10 Griggward Hillslope Models

where colluvial processes create a less stable environment. E horizon depths increase rapidly starting around 75 metres suggesting strong subsurface flow or waterlogging of the surface in the middle of the convergent hillslopes. The E horizons shallow and the A horizons deepen at the convergent hillslope base, perhaps due to linkage with broader catchment alluvial processes.

Inter Study Area Interpretations

Figures 5.11 and 5.12 show that solum depths and standardized hillslope distances decrease and hillslope heights increase from Brucedale south to Ladysmith and Griggward. Comparison of the Ladysmith and Griggward Ordovician metasediment hillslopes shows that a much larger proportion of the solum is A and E horizon in Griggward. These suggest that as climate, basic water balance and terrain change from north to south, biological productivity is increasing as reflected in the regional increase in A horizon total carbon (Table 5.2). The low levels of total carbon at Brucedale may also be due to frequent cereal cropping that reduces organic matter.

The shallower solum depths at Griggward may indicate that lateral surface flow processes are removing materials from the system more so than at Ladysmith. The Ladysmith solum depths deepen considerably from hillslope summit to base perhaps suggesting more local re-distribution where sediments are remaining in the hillslope system.

The greater prevalence of E horizons on both convergent and divergent hillslope models at Griggward may suggest more prevalent lateral subsurface flow, more podzolization, or a combination of the two in comparison to Ladysmith. This hints that the moister, cooler and higher energy Griggward landscapes exhibit a more dynamic mixture of processes including overland flow, material movement by surface erosion and subsurface flow that re-distributes biogeochemical materials (e.g. solutes, colloids, soil aggregates).

Comparison of models in Figures 5.11 and 5.12 suggest that, among other local factors, regional climate is generating soil pattern differences.

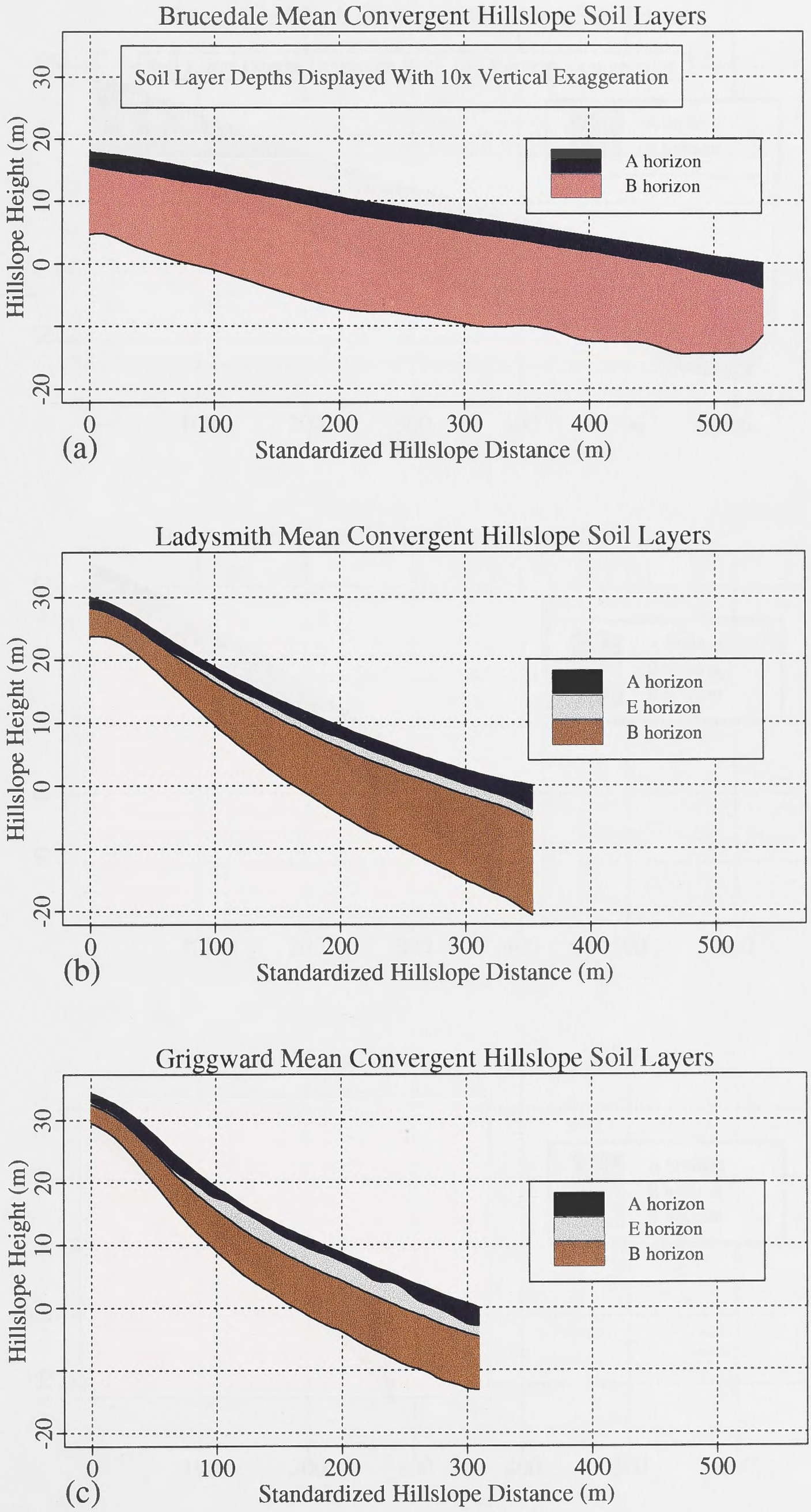
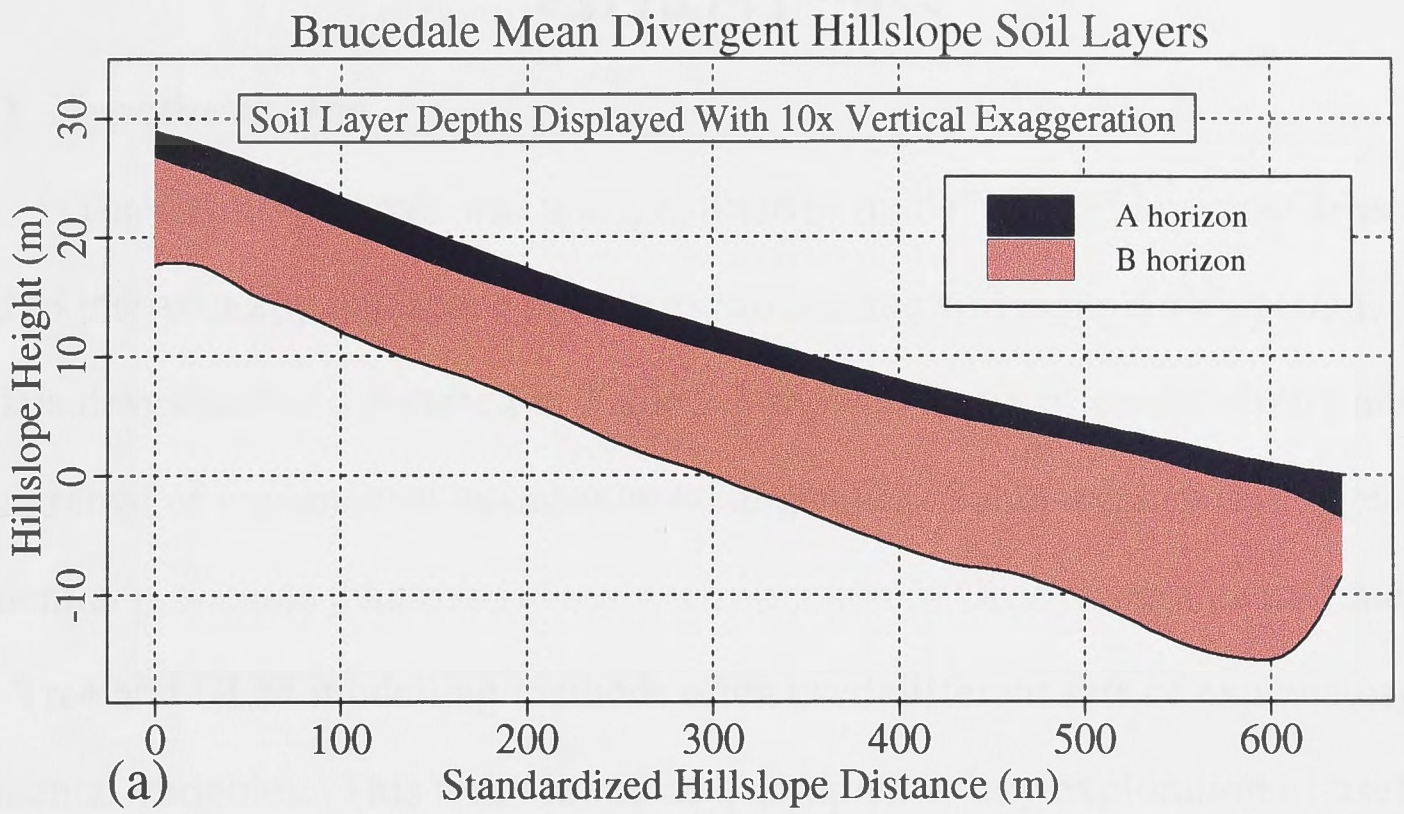
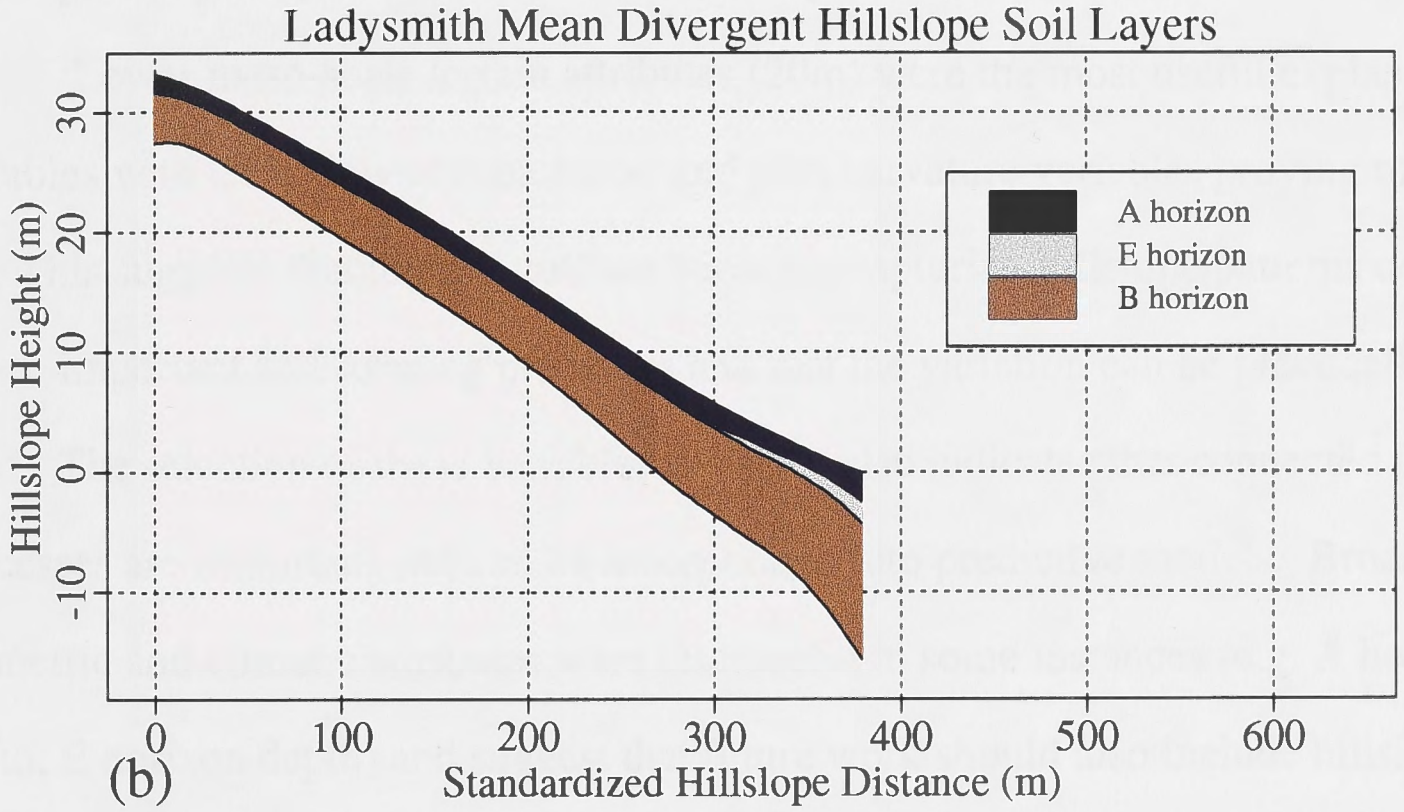


Figure 5.11 Mean Convergent Hillslope Models





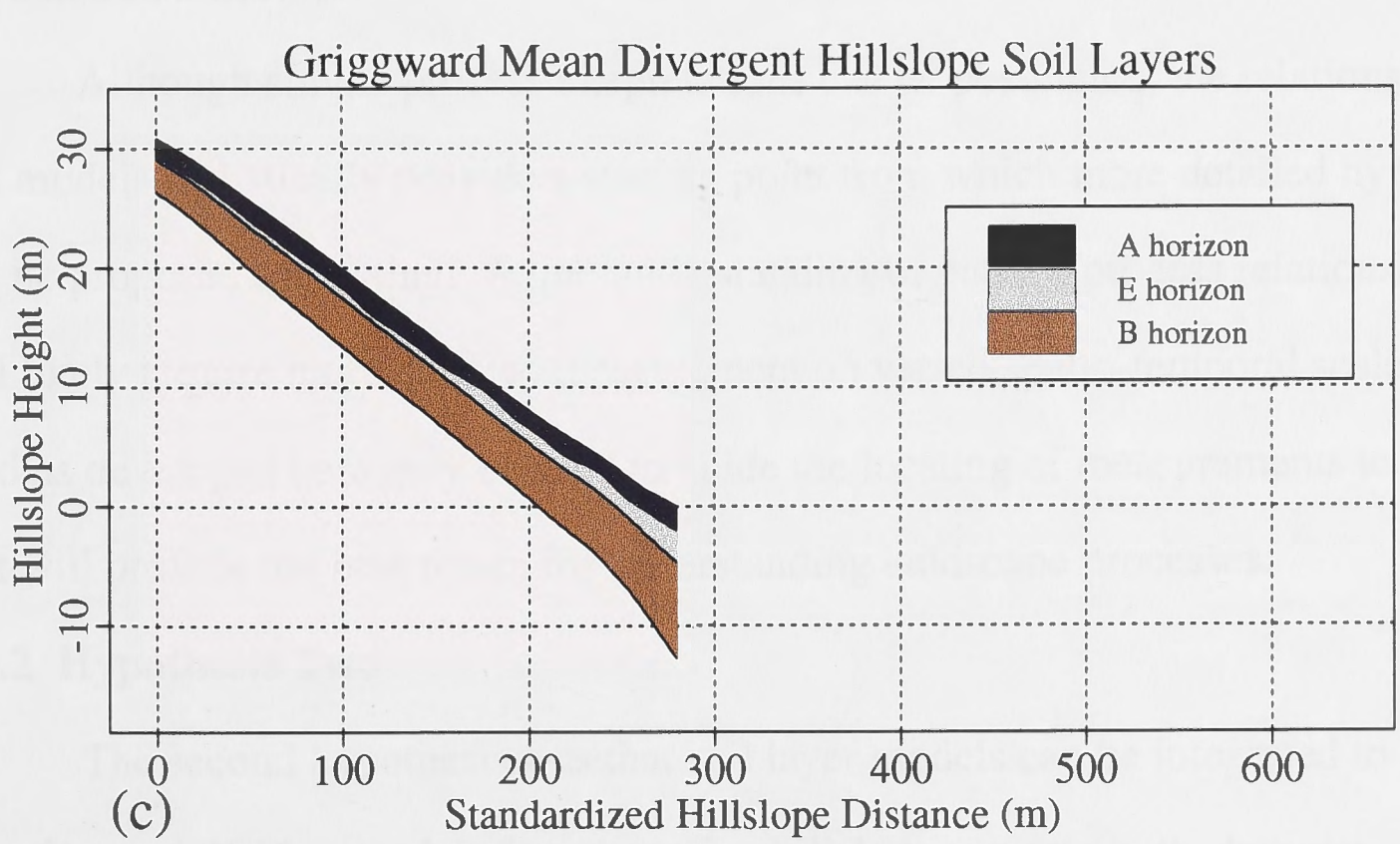


Figure 5.12 Mean Divergent Hillslope Models

5.4 CONCLUSIONS

5.4.1 Hypothesis One

The first hypothesis was that quantitative models of soil layer patterns can be used to interpret key landscape processes influencing soil layer development. The models developed and reported in Table 5.1 showed a mix of predictability and used a broad range of explanatory environmental attributes. Some were more suggestive of influential processes than others (e.g. podzolization in Ordovician metasediments). The Tree and GLM modelling methods often used different sets of explanatory environmental variables. This was viewed as a complementary exploration of useful relationships and perhaps indicates that simplistic process interpretations are unrealistic.

Lower meso-scale terrain attributes (20m) were the most useful explanatory variables with CTI, flow accumulation and plan curvature variables proving most useful. This suggests that terrain attribute variation capturing hillslope patterns does relate to important soil forming processes and that the variation can be practically modelled. The selection of these variables as useful also indicates that connected hillslope processes are important and can be incorporated into predictive models. Broader radiometric and climatic attributes were also useful in some instances (e.g. A horizon depth, E horizon depth) and suggest that future work should also include hillslope solar radiation attributes.

Although simple process intepretations can be postulated, the relationships and models realistically provide a starting point from which more detailed hypotheses can be proposed for testing. A true understanding of pattern/process relationships will likely require more detailed measurement on varied spatio-temporal scales. The models developed here may be used to guide the locating of measurements to areas that will provide the best return for understanding landscape processes.

5.4.2 Hypothesis Two

The second hypothesis was that soil layer models can be integrated to provide quantitative hillslope models for comparing hillslope patterns in the broader

environmental context. A quantitative hillslope model was defined and used to sample and demonstrate the integration of individual soil layer models for development of spatially averaged hillslope models. Important hydrological and geomorphological concepts of hillslope connectivity and adjacency were preserved in models that are representative of patterns over the geographic extent of each study area. The visualizations usefully conveyed differences of the structure of the soil-landscape over the more complex three-dimensional landscape using convergent and divergent end-members as quantified by digital terrain attributes.

Many interpretations can be postulated from the quantitative hillslope visualizations. Some, and perhaps all, may be incorrect. The important advance is that the methods provide a framework from which more informed hypotheses and questions can be posed based on the empirical evidence gathered, analyzed and visualized. Understanding and visualizing the structure of the soil-landscape continuum in a quantitative manner is the first step towards developing an understanding of dynamic soil-landscape function.

Additional comparative analyses that pool the data from the three study areas to build predictive models and perform analysis of deviance will shed more light on the importance of regional climatic factors in interpreting regional soil patterns. This would provide a useful link to the broader environmental context that would complement the lower meso-scale models developed in each study area.

The key findings in this Chapter were:

- quantitative models of soil layer patterns provide a guide to the suite of processes and important scales (e.g. local versus contextual) that influence soil layer patterns, but relationships are complex perhaps due to the varied scales of measurement; and

- a broad array of tools can be integrated to develop advanced visualizations of three dimensional soil layer patterns representative of a spatial area.

5.5 REFERENCES CITED

Bierwirth, P., P.E. Gessler, and D.J. McKane. 1996. Investigation of airborne

gamma-ray images as an indicator of soil properties - Wagga Wagga, NSW. *In* Proceedings of the 8th Australasian Remote Sensing Conference, Canberra. 25-28 March 1996. Canberra.

Bierwirth, P. in preparation. Investigation of airborne gamma-ray images as a rapid mapping tool for soil and land degradation: Wagga Wagga. AGSO Record Interp., Australian Geological Survey Organization. Canberra, Australia.

Buol, S.W., F.D. Hole, and R.J. McCracken. 1989. Soil genesis and classification. 3rd Edition. Iowa State Univ. Press, Ames, IA.

Butler, B.E. 1956. Parna: an aeolian clay. Aust. J. Sci. 18:145-151.

Butler, B.E. 1959. Periodic phenomena in landscapes as a basis for soil studies. CSIRO Australia. Soil Publ. No. 14. Adelaide, Australia.

Carson, M.A., and J.J. Kirkby. 1972. Hillslope form and processes. Cambridge University Press, Cambridge.

Daniels, R.B., and R.D. Hammer. 1992. Soil geomorphology. Wiley and Sons, New York.

Forman, R.T.T., and M. Godron. 1986. Landscape ecology. Wiley and Sons, New York.

Gerrard, A.J. 1981. Soils and landforms. Allen & Unwin, London.

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modelling and spatial prediction of soil attributes. Int. J. Geographical Information Systems, Vol. 9, 4:421-432.

Hole, F.D., and J.B. Campbell. 1985. Soil landscape analysis. Rowman & Allenheld, Totowa.

Jenny, H. 1941. Factors of soil formation: a system of quantitative pedology. McGraw Hill, New York.

Jenny, H. 1980. The soil resource: origin and behavior. Ecological Studies 37. Springer-Verlag, New York.

McMahon, J.P., M.F. Hutchinson, H.A. Nix, and K.D. Ord. 1995. ANUCLIM: users guide. Centre for Resource & Environmental Studies, Australian National University. Canberra, Australia.

McSweeney, K., P.E. Gessler, B. Slater, R.D. Hammer, J. Bell, and G.W. Petersen. 1994. Towards a new framework for modelling the soil-landscape continuum. p.127-145. *In* Factors of soil formation: a fiftieth anniversary retrospective. SSSA Special Pub. 33. Madison, WI.

Milne, G. 1935. Some suggested units of classification and mapping particularly for East African soils. Soil Research 4:3.

Moore, I.D., A.R. Ladson, and R. Grayson. 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydro. Proc. 5:3-30.

Moore, I.D., P.E. Gessler, G.A. Neilsen, and G.A. Peterson. 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57:443-452.

Naveh, Z., and A.S. Lieberman. 1984. Landscape ecology: theory and application. Springer-Verlag, New York.

Raymond, O.L. 1992. The geology of Wagga Wagga and the Kyeamba Valley. 1:100,000 scale preliminary edition. Australian Geological Survey Organisation.

Ruhe, R.V. 1975. Geomorphology. Houghton Mifflin, Boston.

Selby, M.J. 1982. Hillslope materials and processes. Oxford Univ. Press, Oxford.

Simonson, R.W. 1959. Outline of a generalized theory of soil genesis. Soil Sci. Soc. of Am. Proc. 23:152-156.

Speight, J.G. 1977. Land form pattern description from aerial photographs. Photogrammetrica. 32:161-182.

Statistical Sciences. 1993. S-PLUS Guide to statistical and mathematical analysis. Version 3.2. StatSci, a Division of MathSoft, Inc., Seattle, WA.

Turner, M.G., and R.H. Gardner. 1991. Quantitative methods in landscape ecology: an introduction. p3-14. *In* Quantitative methods in landscape ecology. Ecological Studies 82. Springer-Verlag, New York.

Walker, P.A., and Butler, B. 1983. Fluvial processes. p.83-91. *In* Soils: and Australian viewpoint. Division of Soils, CSIRO. Academic, London.

Chapter Six: Conclusions and Recommendations

This thesis demonstrates an integration of tools and methods for development of statistical soil-landscape models. The models are based on explicit and quantitatively defined environmental correlations between soil attributes and a broad range of environmental variables. This approach has many parrallels with traditional soil survey in that samples of a population are used to develop relationships for spatial extension. However, in this approach spatial analysis tools are used with digital data to explicitly define the study areas at the upper meso-scale (the level above the intended application) and develop a provisional model for a statistically-based stratified random sampling along a CTI environmental gradient at the hillslope level (lower meso-scale). This incorporates information about the physical characteristics of the landscapes under study and, I postulate, improves the quality of spatial prediction and potential for hypothesizing soil-landscape processes based on quantitative sample evidence.

Chapters One and Two established the conceptual framework and reviewed literature for the statistical soil-landscape modelling approach used here. Six thesis hypotheses were tested using soil and related environmental data measured and collected at various scales over three study areas.

6.1 HYPOTHESIS ONE

The first hypothesis was that explicit and quantitative environmental correlations can be derived to spatially predict individual soil attributes using statistical models with stated levels of uncertainty and model complexity. Chapter Three confirmed the hypothesis. Statistical sampling, exploratory data analysis and confirmatory statistical modelling were used to develop mathematical relationships for spatial implementation and visualization. Models for solum depth, total carbon, cation exchange capacity and exchangeable sodium percentage show that a flexible analysis approach using iterative EDA and modelling tools (GLM, GAM, Tree) is required to search for,

and evaluate useful environmental correlations. Varied approaches are required because the data exhibit different patterns of variation down hillslopes and through the soil profile.

Solum depth exhibits a strong relationship with lower meso-scale (20m) terrain attributes and is predicted with a high level of certainty (%RID =78). Total carbon exhibits a strong and very smooth relationship with soil depth but not with landscape and is also predicted with a high level of certainty (%RID = 84). Cation exchange capacity variation is significantly partitioned by horizon, but the variation is still widely scattered and not as predictable. A B-horizon model of CEC was illustrated with a %RID of 21. Exchangeable sodium percentage values are very low overall (%RID 11, B horizon) but show a geographic clustering of high values in low, level landscape positions.

An intregrated model using solum depth, A horizon depth, a scatterplot smoother GAM of total carbon and assumptions of horizon bulk densities is implemented to demonstrate how models may be combined to model variation through the soil profile over the landscape. Although the ESP is predicted with a very low level of certainty, simple spatial rules were derived from a regression Tree model to produce a simplistic B horizon sodicity risk map. Visualization of developed models using standard GIS tools for generating colour maps and drapes over the landscape concisely illustrate the level of certainty of each model and level of resolution.

It is concluded that the approach demonstrated in Chapter Three is feasible and holds great potential for implementation of explicit and quantitative statistical soil-landscape models using the techniques, widely available environmental variables (e.g. 20 digital terrain attributes) and integrated statistical and GIS modelling tools. Further recommendations are provided below.

6.2 HYPOTHESIS TWO

The second hypothesis was that quantitative terrain attributes change systematically with scale. Chapter Four demonstrates that this hypothesis is accepted, based on the

methods tested in the small Ordovician metasediment study area. Q-Q plots are a useful tool for summarizing a large amount of data to systematically visualize changes in terrain distribution with varied grid point spacing. The analysis suggests that quantitative scaling equations may be feasible to move from small grid spacings to larger grid spacings. Understanding these changes should assist in research evaluating scaling in various directions up and down the space-time continuum.

6.3 HYPOTHESIS THREE

The third hypothesis was that certain terrain attribute grid point resolutions exist where soil attribute prediction is better (as measured by %RID). Chapter Four demonstrates that this hypothesis is rejected based on the analysis and sample evidence used here. Several grid point resolutions exhibited useful correlations with the soil layer attributes as measured from soil cores. However, in some instances there are resolutions that appear optimal for individual soil attributes. A definite decrease in predictive capacity is seen at grid point spacings beyond 40m (e.g. 80m) suggesting the loss of important landscape variation at 80m spacing in the Ordovician metasediment landscapes. These results are likely specific to this landscape and physiographic domain. The analysis does not indicate that the 20m grid spacing, used more broadly in this work, is a poor grid spacing for developing useful terrain attribute environmental correlations.

6.4 HYPOTHESIS FOUR

The fourth hypothesis was that more detailed topographic data sources, closer to the scale of the soil attribute sample measurements, provide better predictions. This hypothesis is rejected by the results in Chapter Four, but further cost-benefit analysis is required. Two topographic data sources were compared (e.g. 1:25k, 1:10k) and terrain attributes from the 1:10 000 source generally provide better predictions of the soil layer attributes. However, collection of topographic data at this scale for broader regional soil-landscape modelling and soil survey would be costly. The differences in predictive capacity with terrain attributes derived from the widely

available 1:25 000 are not large. The results suggest that the extra data collected as spot heights along ridge-tops and streamlines may be the key difference between the two sources. These types of data could be easily generated from various sources (e.g. digitized from topographic maps, GPS field collection) to supplement the standard 1:25 000 data.

6.5 HYPOTHESIS FIVE

The fifth hypothesis was that environmental correlations defined by quantitative models of soil layer patterns can be used to interpret key landscape processes influencing soil layer development. Chapter Five supports this hypothesis. This conclusion is based on the premise that soil data are typically very noisy and difficult to predict in the broader spatial context important for soil survey . This is coupled with the fact that measurement of soil response and environmental predictor variables is limited to finite increments over the space-time continuum (i.e. processes are likely multi-scaled and dynamic). Although some of the soil layer models presented in Chapter Five do not have obvious process interpretations, they provide a basis for improved hypotheses on processes and these can be further tested and refined. The modelling approach aims to collect data to allow re-analysis and improvement of models and understanding over time. Explicit and quantitative methods leading to collation of data in a GIS provide a framework that facilitates continued analysis and communication with other disciplines. This point is expanded below.

6.6 HYPOTHESIS SIX

The sixth hypothesis was that developed soil layer models can be integrated to provide spatially averaged and quantitative hillslope models to compare and contrast the structure and develop hypotheses about function of soil-landscapes in the broader environmental context of the three study areas. Chapter Five supports this hypothesis. Digital terrain analysis tools were integrated for modelling hillslope connectivity and adjacency for use in sampling soil layer models over the three study areas. The result is spatially-averaged and quantitative hillslope models representing

convergent and divergent portions of the three-dimensional landscape. Presentation as cross-section displays of the soil layer patterns efficiently communicates an integrated view of the soil-landscape structure that may improve the development of hypotheses regarding landscape function. Visualizations of the soil layer patterns on a single graphic highlights important differences in soil patterns over the broader spatial region containing the study areas and suggests that climatic differences may be present.

Although spatially averaged representations of soil layer patterns are presented, any other environmental variable available in a spatially continuous manner over the study areas can be summarized using the methods demonstrated. The method is useful for visualizing integrated data or models along hypothesized flow vectors and pathways of material and energy on hillslopes. It should assist in communicating and linking models of the soil-landscape continuum into broader ecological and socio-economic models that require the development of a more integrated understanding of landscapes.

6.7 RECOMMENDATIONS

Some recommendations can be made with the advantage of hindsight. If this research was repeated with the intention of developing soil-landscape models at the lower meso-scale, several recommendations can be made. First, development of upper meso-scale variables (e.g. geology, climate surfaces) should proceed well before field sampling to ensure that ample time is allotted for stratification and selection of representative catchments. The development of the broad range of geographic datasets used here by necessity occurred in concert.

Second, a controlling factor in the spatial extent of the three study areas was GIS disk space limitations. This prevented a broader analysis of lower-mesoscale terrain attributes prior to study area selection. Following this, use of variograms to quantify the spatial scale of variation of an environmental variable provides useful information, but does not allow incorporation of process understanding into the statistic.

A useful supplement to the variogram that computes semivariance in all directions would be a variogram that computes semivariance along hypothesized hydrological flow vectors. The usefulness of flow accumulation and related specific catchment area verifies that connectivity, adjacency and accumulation are important parameters. Development of spatial statistics that incorporate these would be useful.

The tools and techniques integrated here are fully within the grasp of any agency or group conducting soil, land and environmental resource inventory, monitoring and management. To incorporate these procedures into a broader, routine survey, several modificaitons may be suggested. First, once a provisional model has been chosen for sampling, the procedure for selecting sample points should be automated and ported, perhaps to a laptop so that sample points can be rejected and re-selected in the field. The process should also proceed in an iterative fashion with data collection and data exploration over a broader region to sequentially test potential upper meso-scale stratifications for appropriateness.

These methods would work best if a team approach were used with a minimum of a field pedologist and spatial analyst working together. It may be possible for a properly trained spatial analyst to work in parrallel with several different field pedologists or survey teams if adequate support and infra-structure were established to routinely collate digital data for GIS database construction. Furthermore, such an infrastructure should become part of a broader scheme for systematic development development of digital data for entire states or regions. Many of the techniques outlined here could form the basis for a proto-type expert system to assist with survey implementation.

6.8 CONCLUDING REMARKS

In 1964, Butler (1964) posed the question: "Can pedology be rationalized?" He posed this question as a challenge and warned that pedology was at risk of becoming an isolated discipline classed as a simple descriptive science that uses an isolated and esoteric language to communicate within. Although this thesis, in several

instances, criticizes taxonomic and soil map unit paradigms, it is appreciated that they were developed in an era when digital spatial analysis was not possible. This paradigm has been useful as evidenced by the fact that the methods and principles are still widely used. However, many pedologists view GIS and spatial analysis tools as nothing more than better and more rapid cartographic automation. This is an unnecessarily restricted view.

The approach developed and tested here integrates a broad array of newer tools and demonstrates how spatial analysis and quantitative techniques can be used for modelling soil spatial patterns using environmental correlations that have traditionally been an important part of soil mapping. Quantitative methods enable more appropriate modelling of soil-landscape continuum patterns and assist the development of integrated theories through facilitating simpler communication and interaction with other complementary disciplines (hydrology, geomorphology, geology, ecology, climatology, biology, economics). Butler's (1964) challenge to rationalize pedology still stands, but the tools and methods demonstrated here may place it on firmer scientific footing (Hewitt, 1993) that encourage more sure-footed movements up and down the stairs between scientific disciplines.

'Habit is habit, and not to be flung out of the window,
but coaxed downstairs a step at a time'

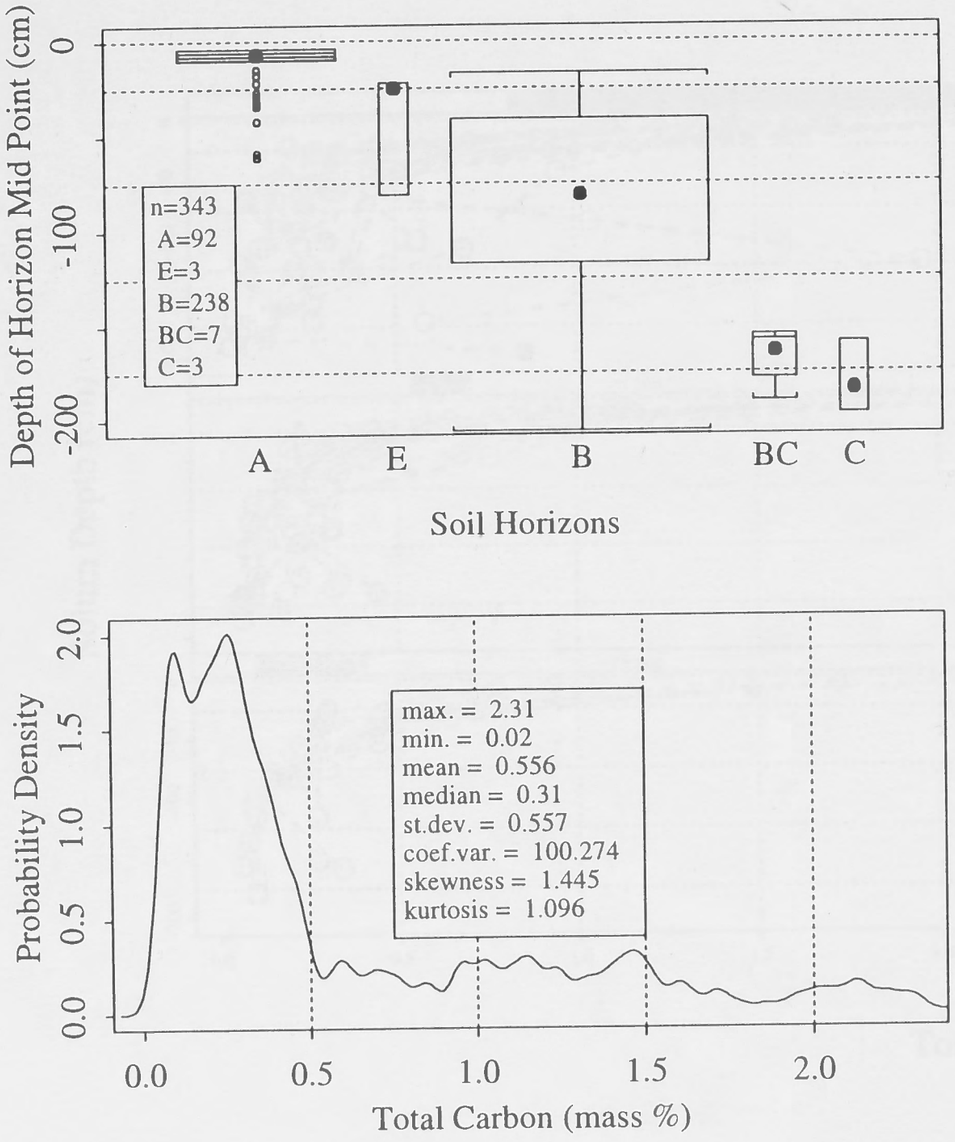
Mark Twain

6.9 REFERENCES CITED

Butler, B.E. 1964. Can pedology be rationalized? A review of the general study of soils. Australian Soc. Soil Sci. Publication No. 3. Canberra, Australia.

Hewitt, A.E. 1993. Predictive modelling in soil survey. Soils Fert. 3:305-314.

Appendix One - Exploratory Data Analysis Graphics



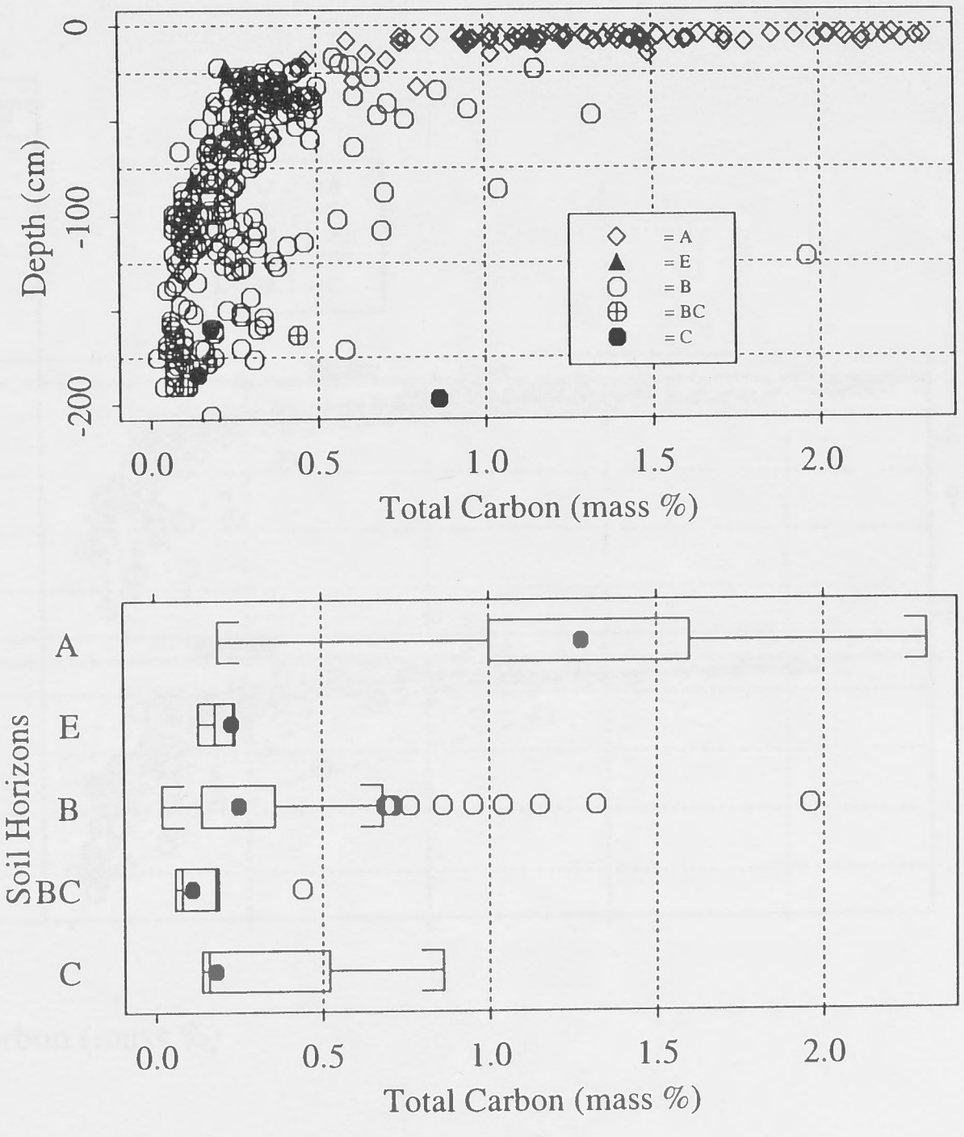


Figure A-1 Total Carbon Univariate and Bivariate EDA (Brucedale)

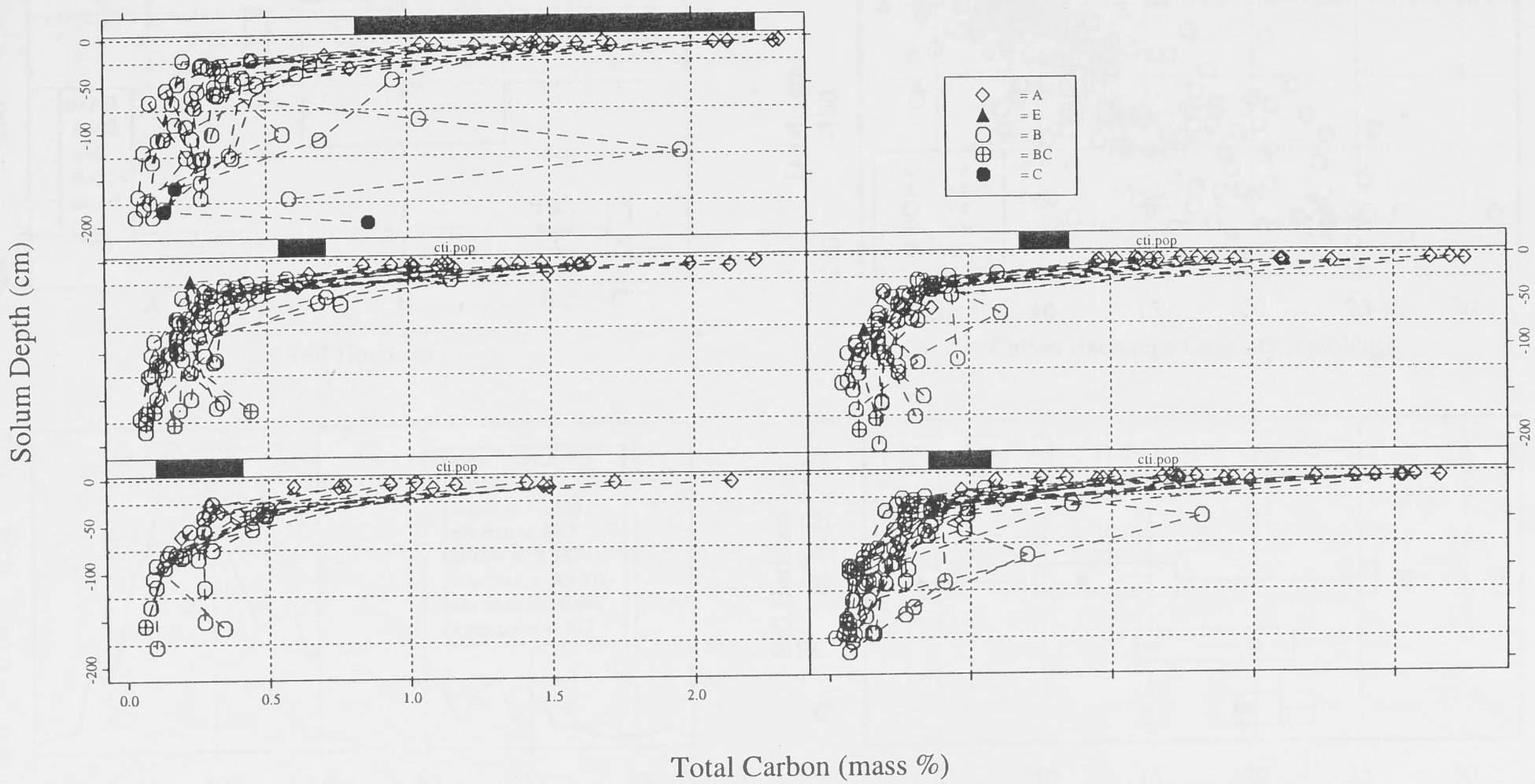
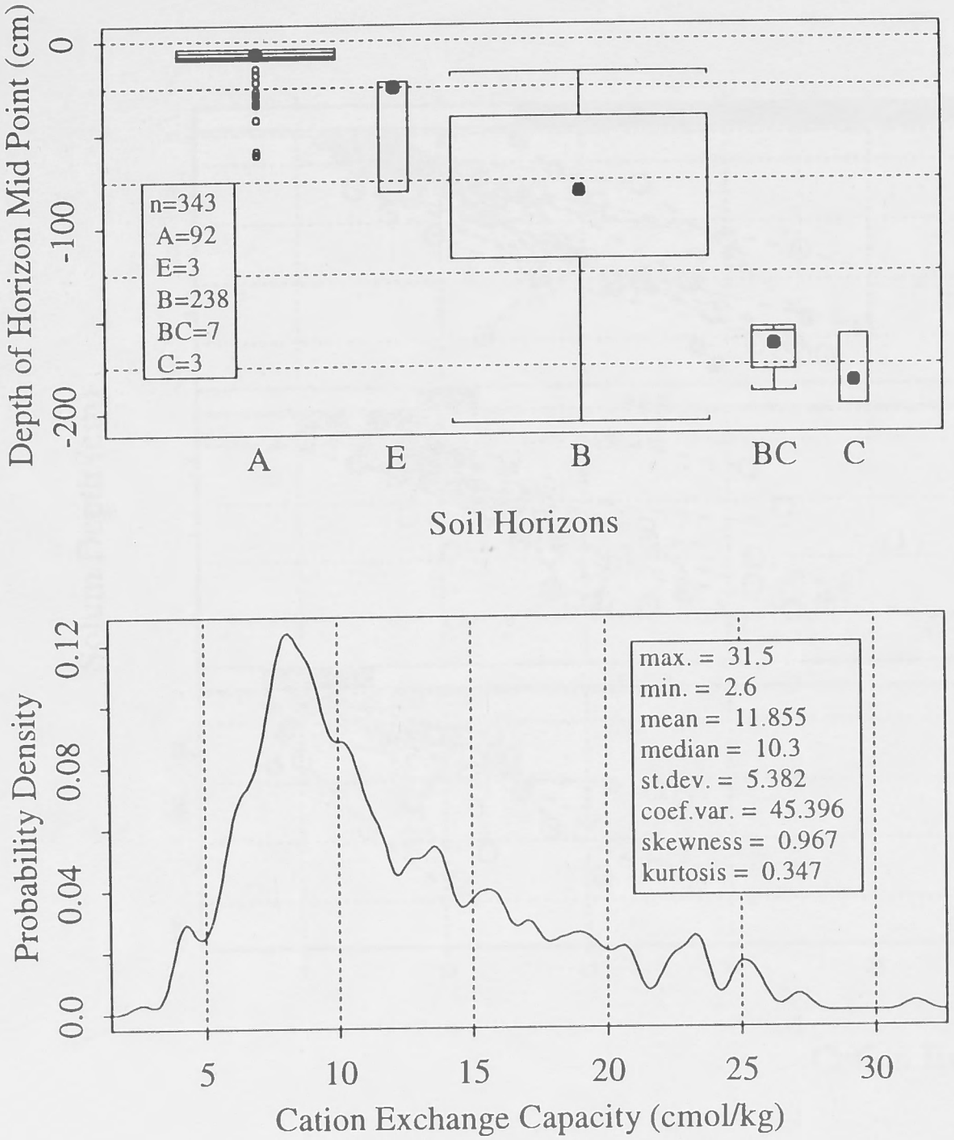


Figure A-2 Total Carbon Trellis Conditioned by Compound Topographic Index (Brucedale)



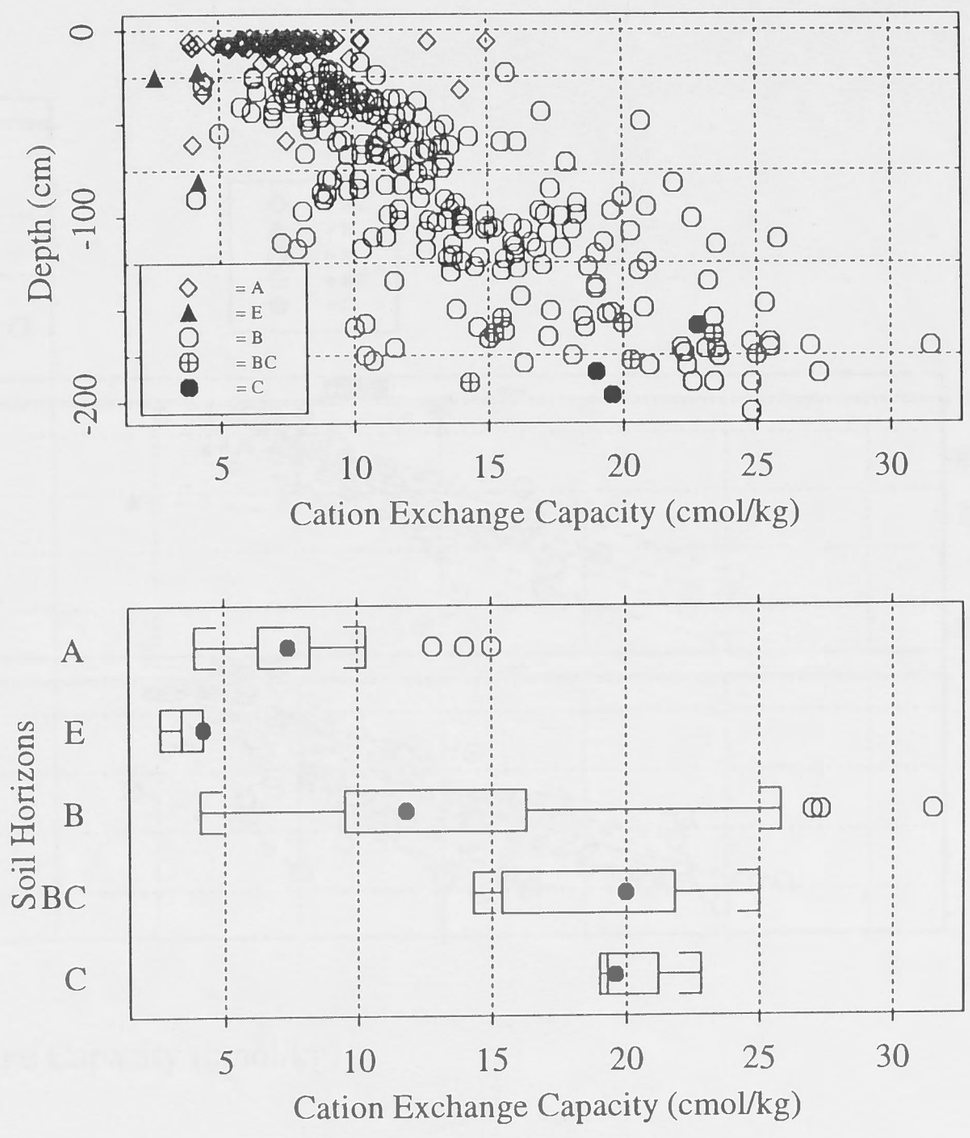


Figure A-3 Cation Exchange Capacity Univariate and Bivariate EDA (Brucedale)

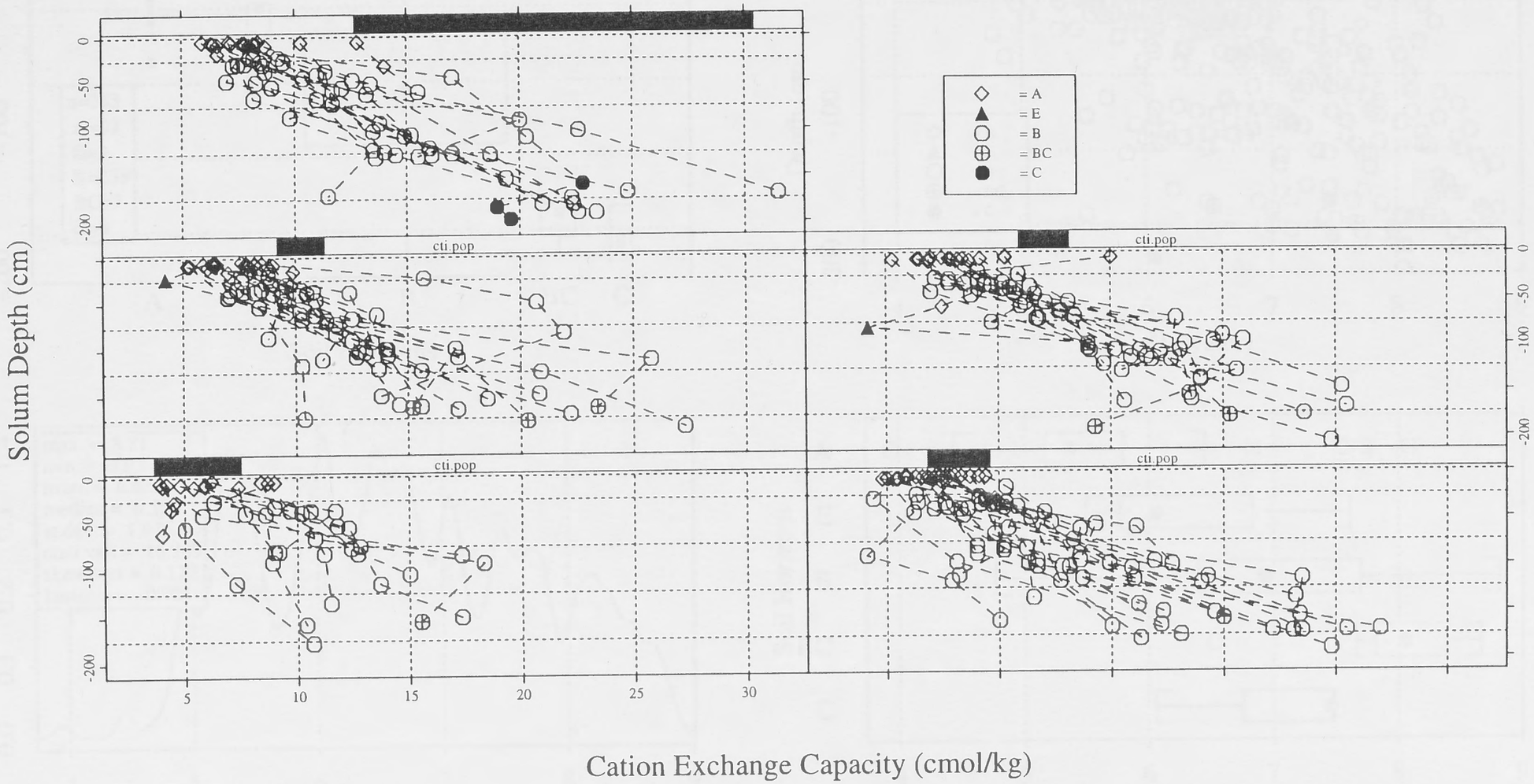
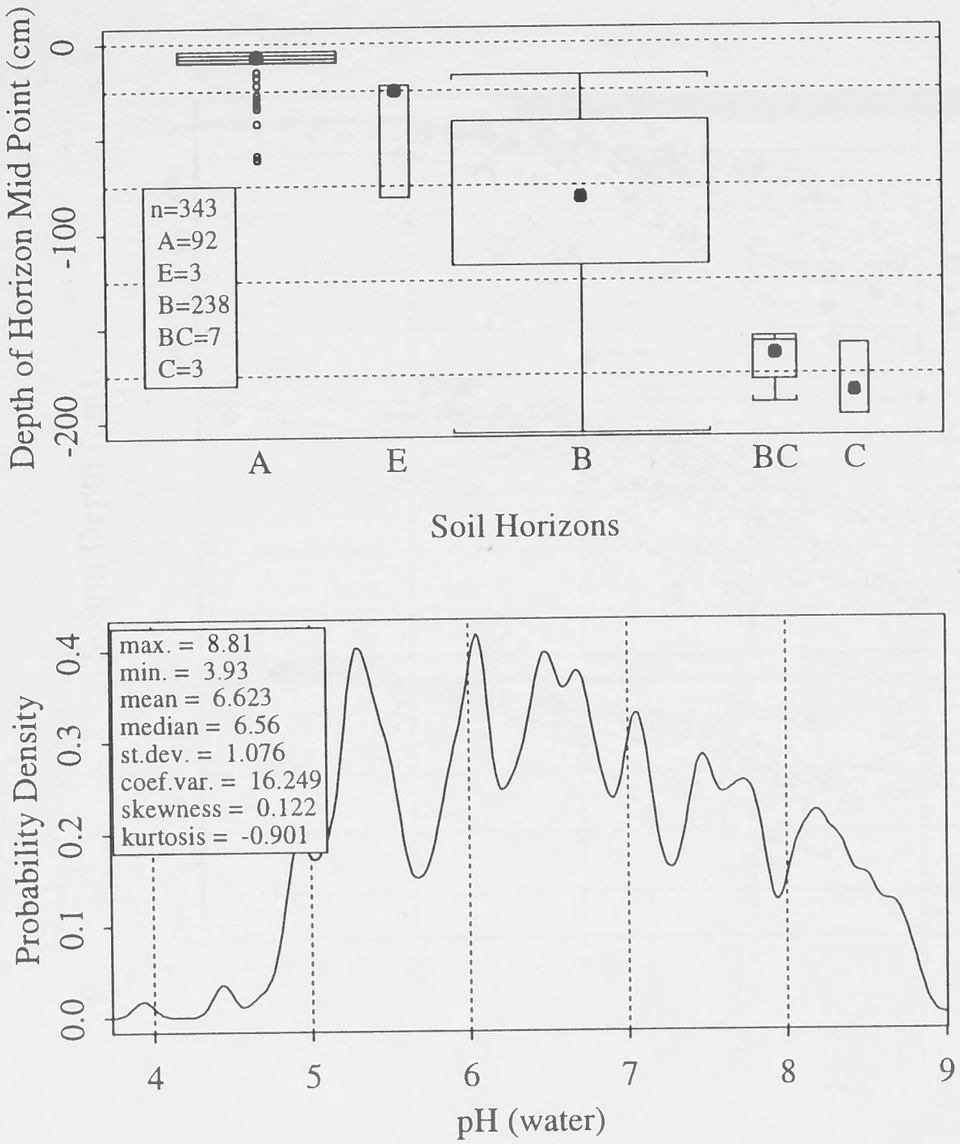


Figure A-4 Cation Exchange Capacity Trellis Conditioned by Compound Topographic Index (Brucedale)



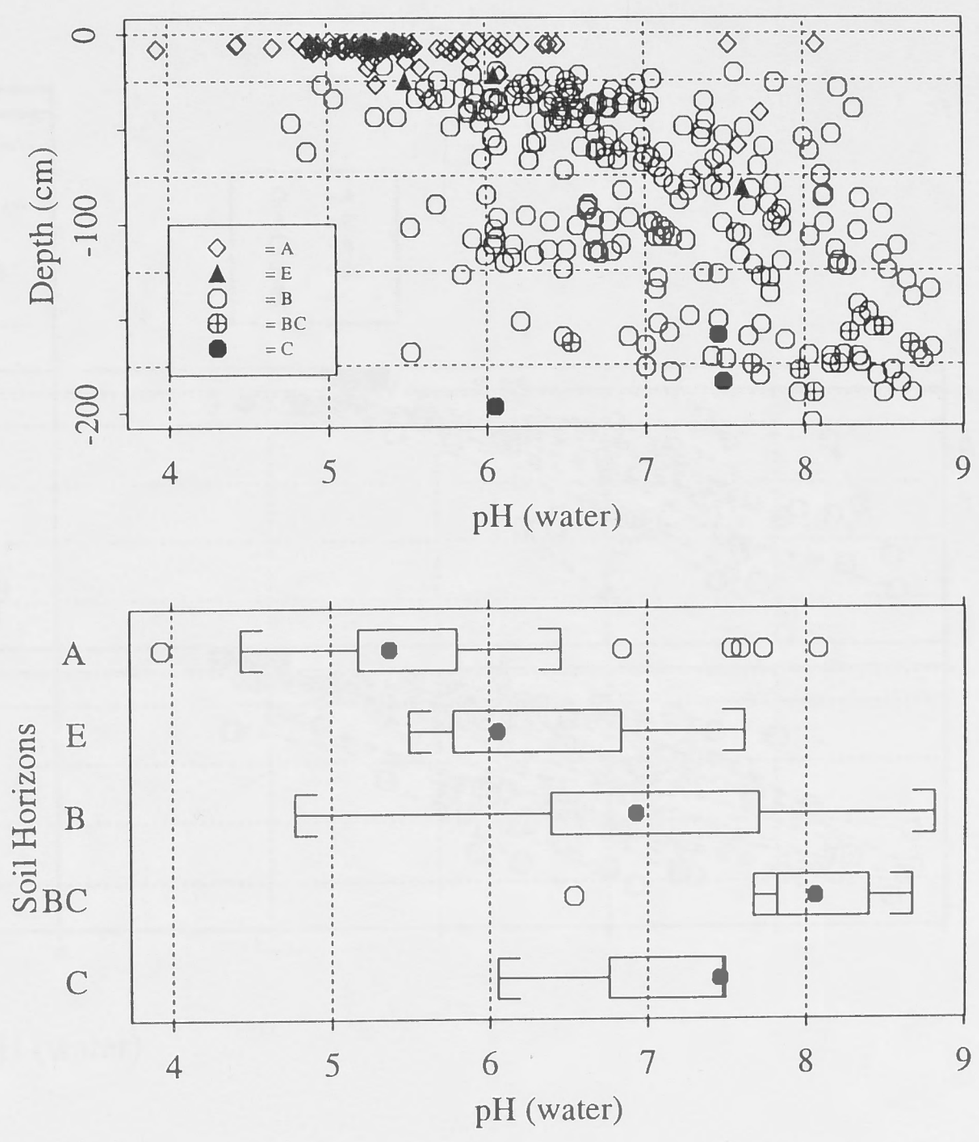


Figure A-5 pH Univariate and Bivariate EDA (Brucedale)

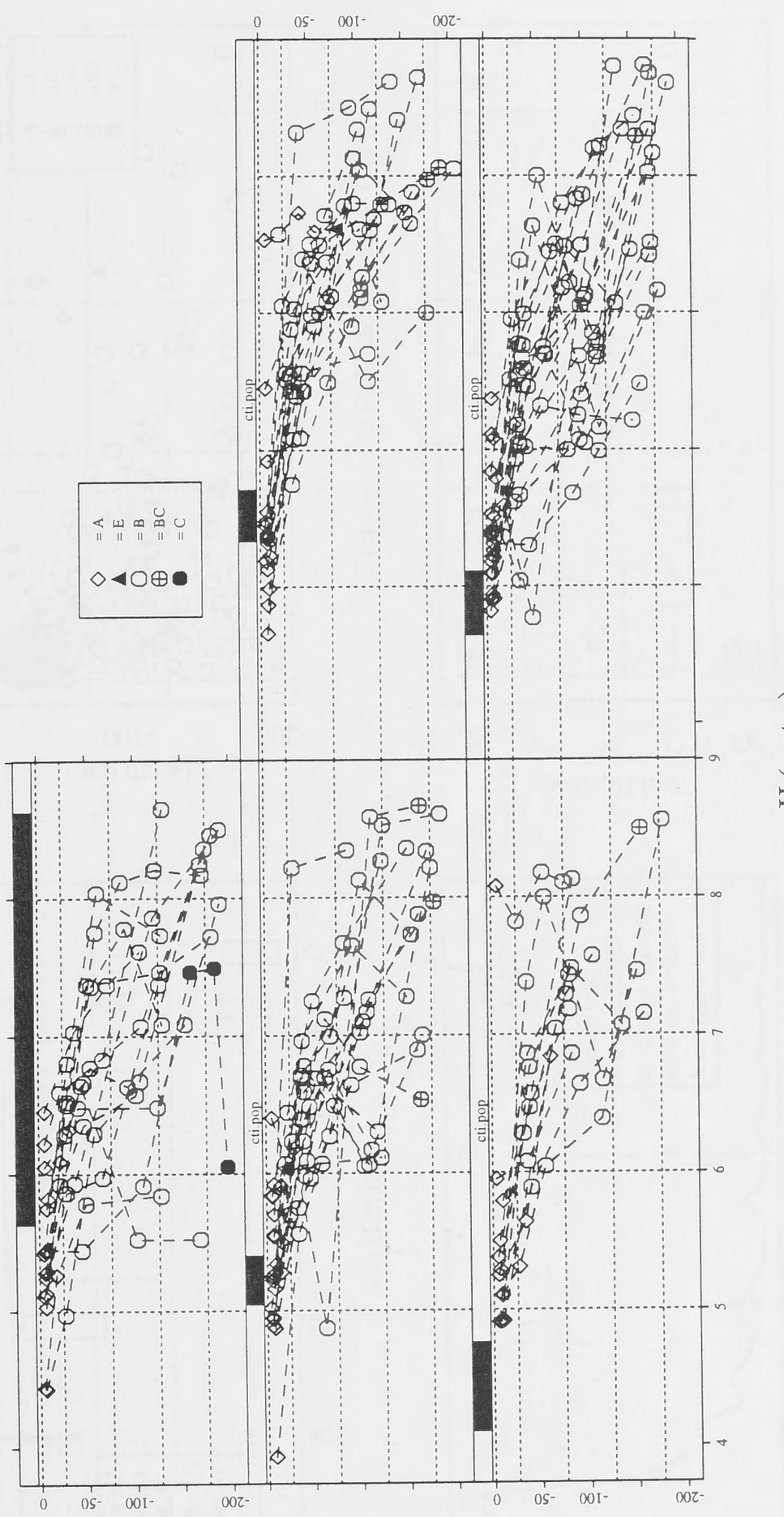
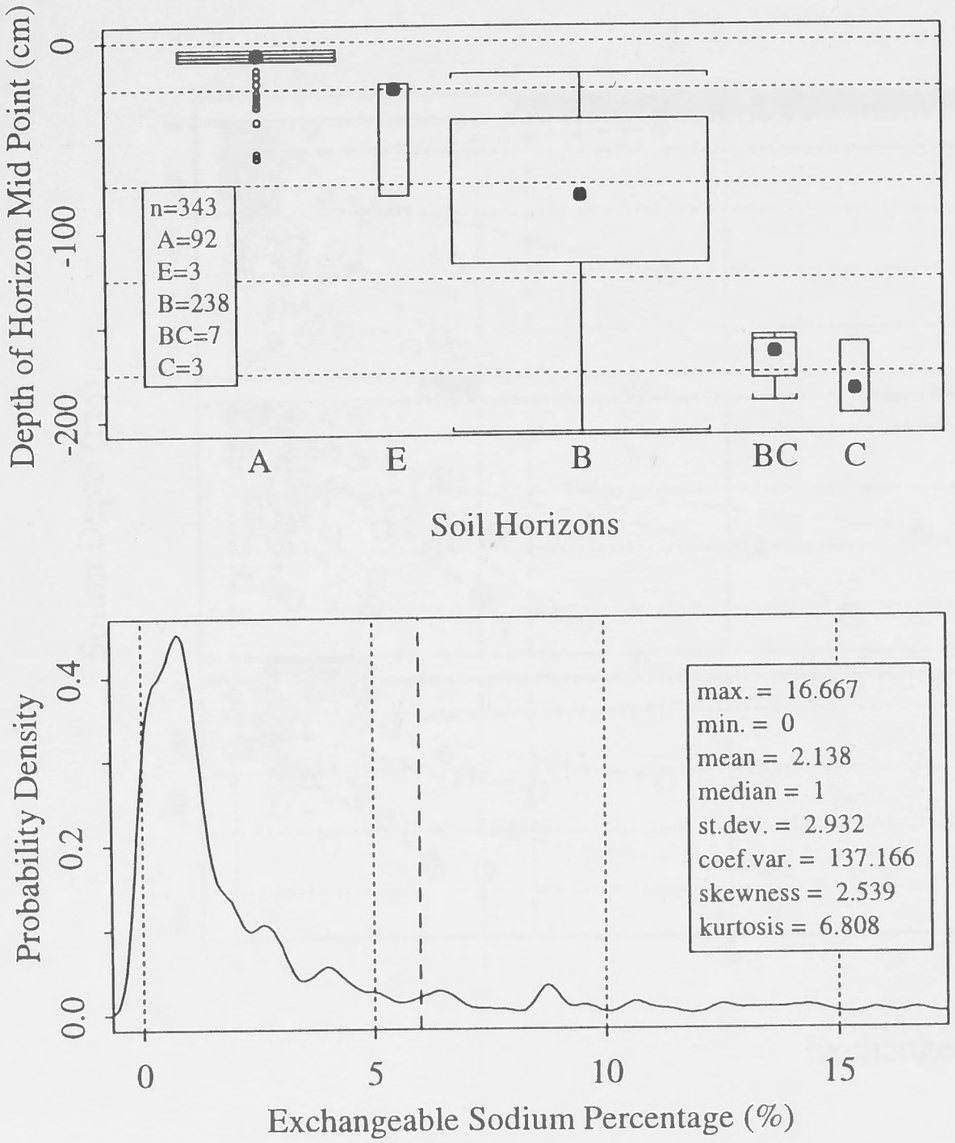


Figure A-6 pH Trellis Conditioned by Compound Topographic Index (Brucedale)



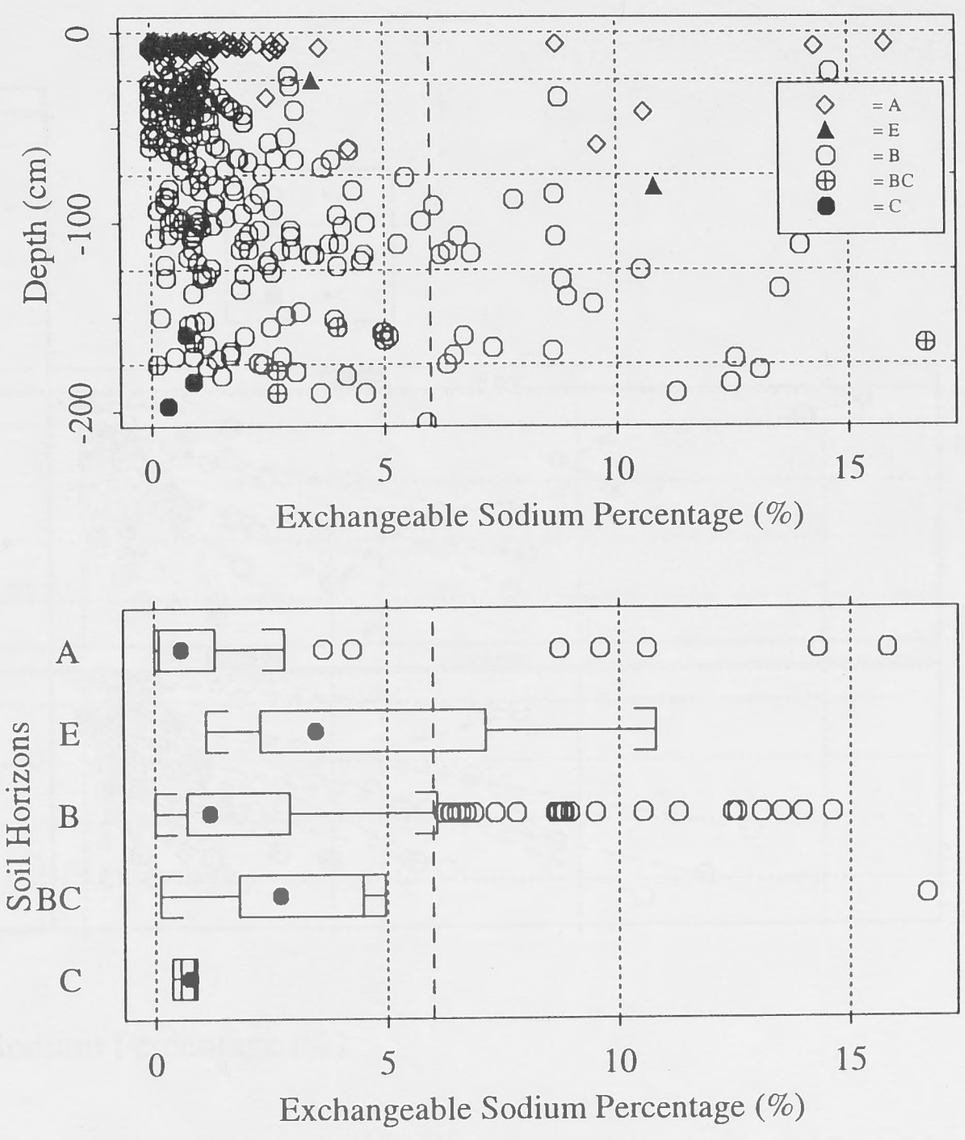


Figure A-7 Exchangeable Sodium Percentage Univariate and Bivariate EDA (Brucedale)

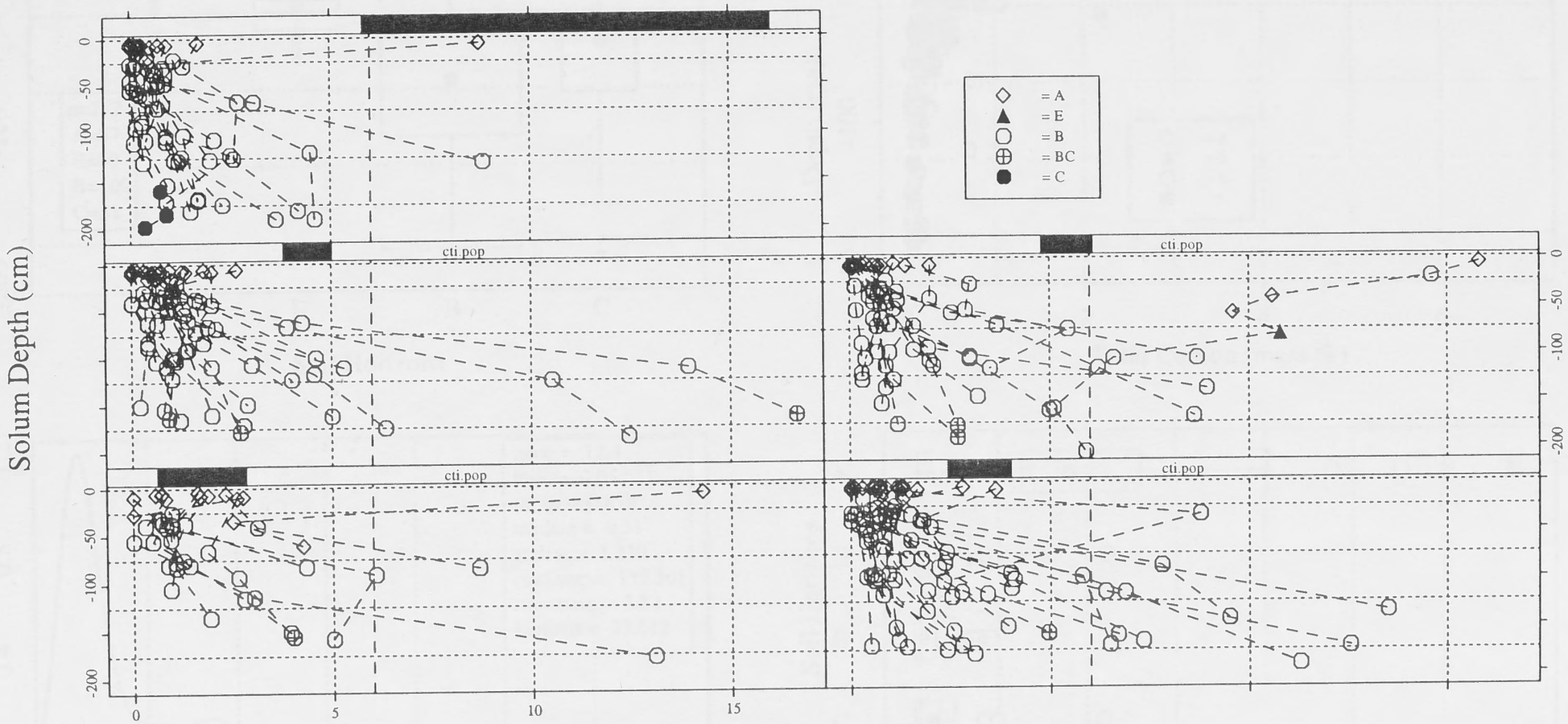
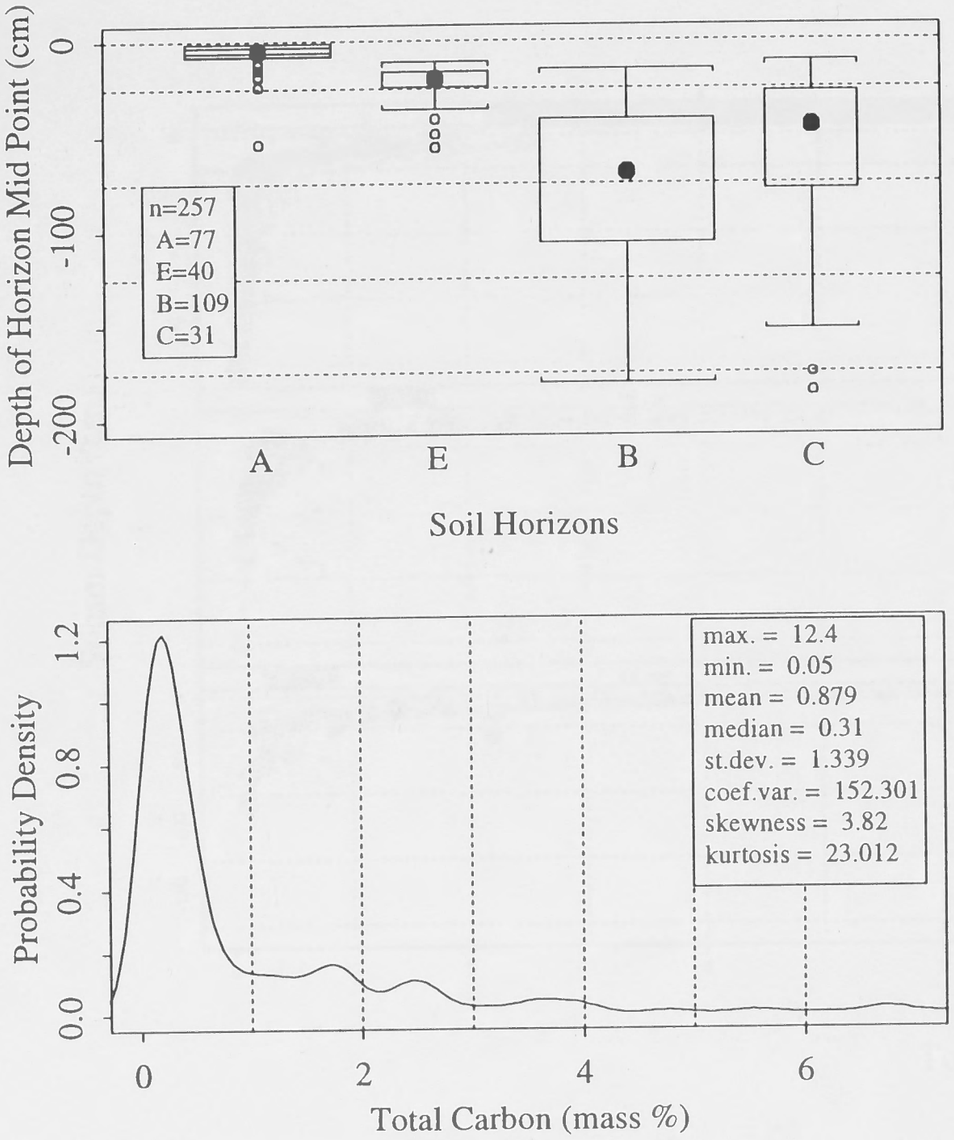


Figure A-8 Exchangeable Sodium Percentage Trellis Conditioned by Compound Topographic Index (Brucedale)



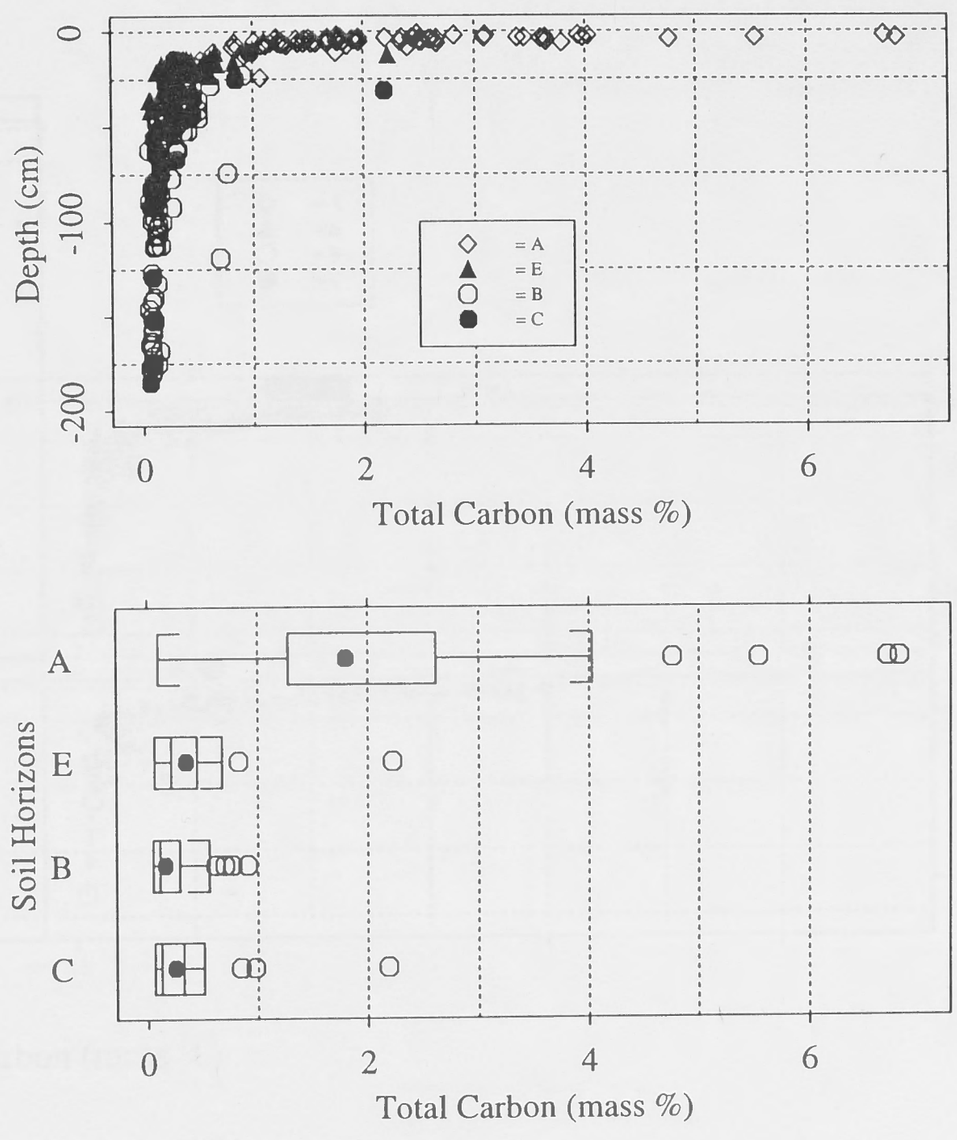


Figure A-9 Total Carbon Univariate and Bivariate EDA (Ladysmith)

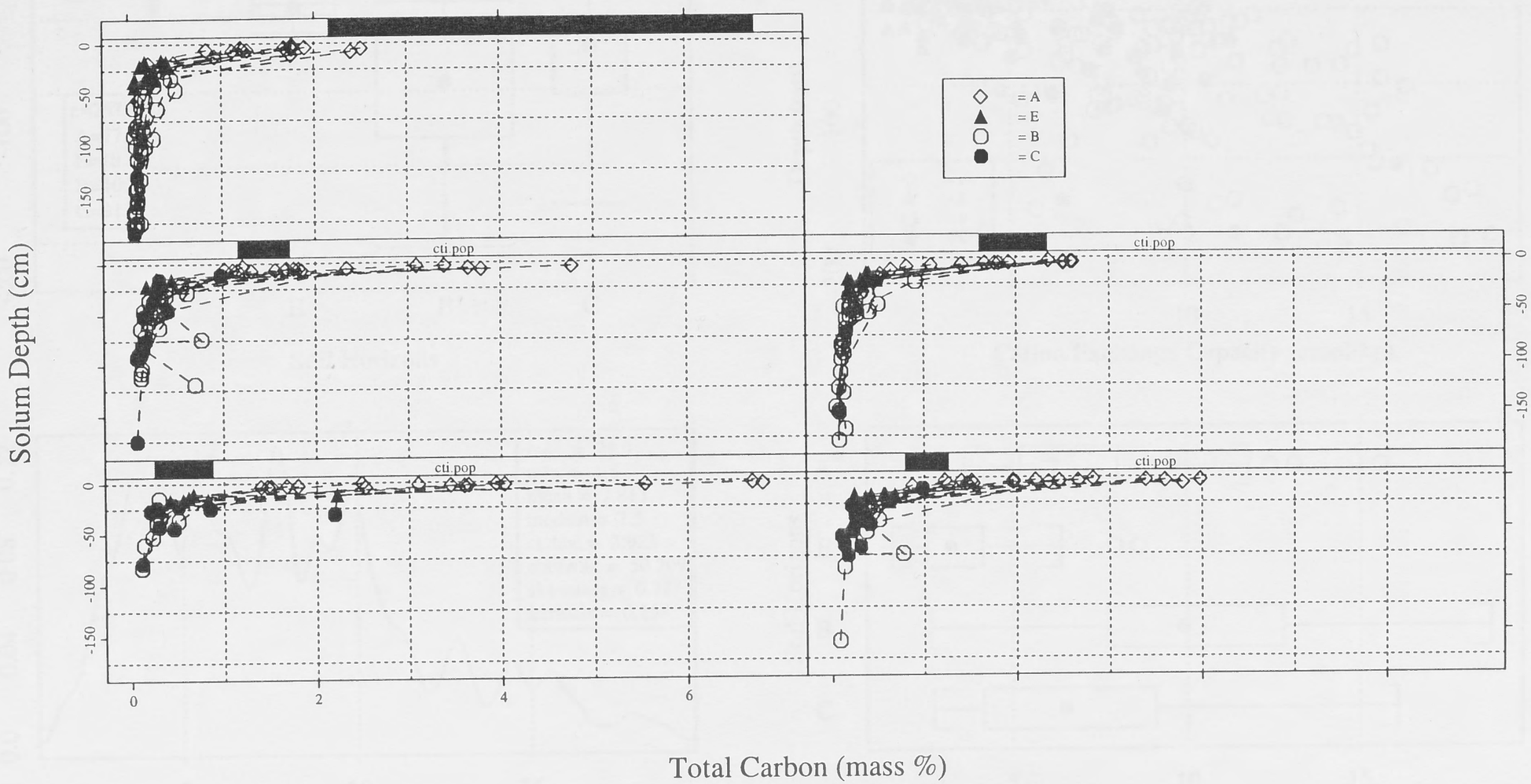
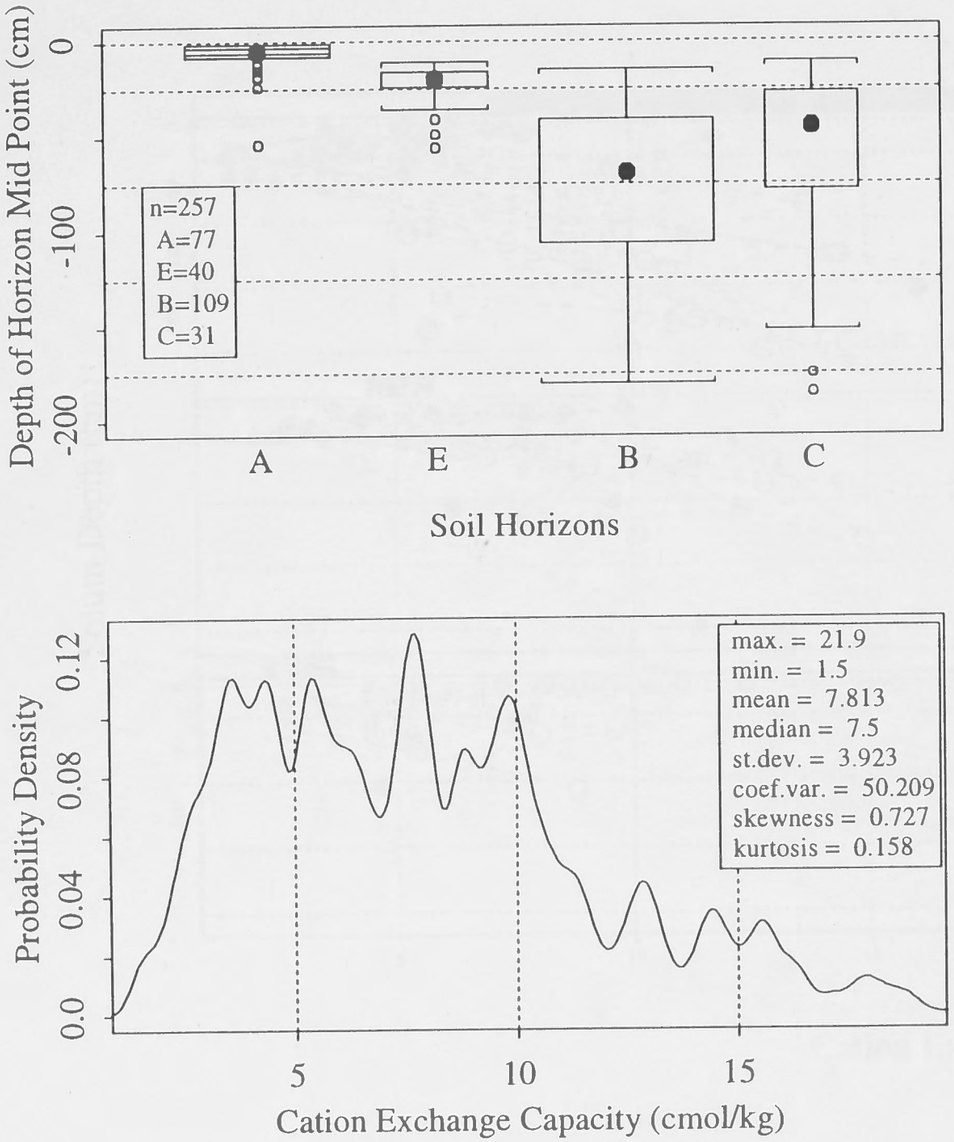


Figure A-10 Total Carbon Trellis Conditioned by Compound Topographic Index (Ladysmith)



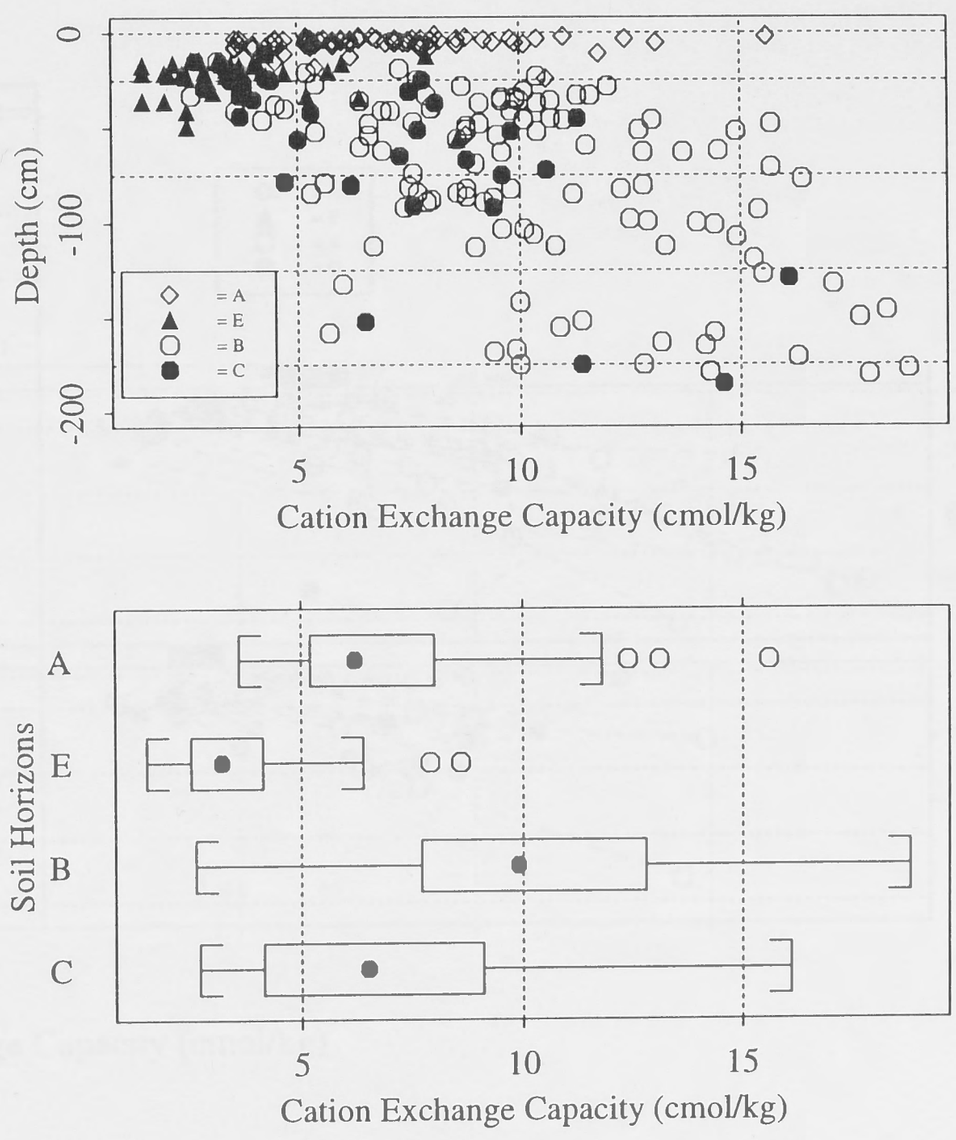


Figure A-11 Cation Exchange Capacity Univariate and Bivariate EDA (Ladysmith)

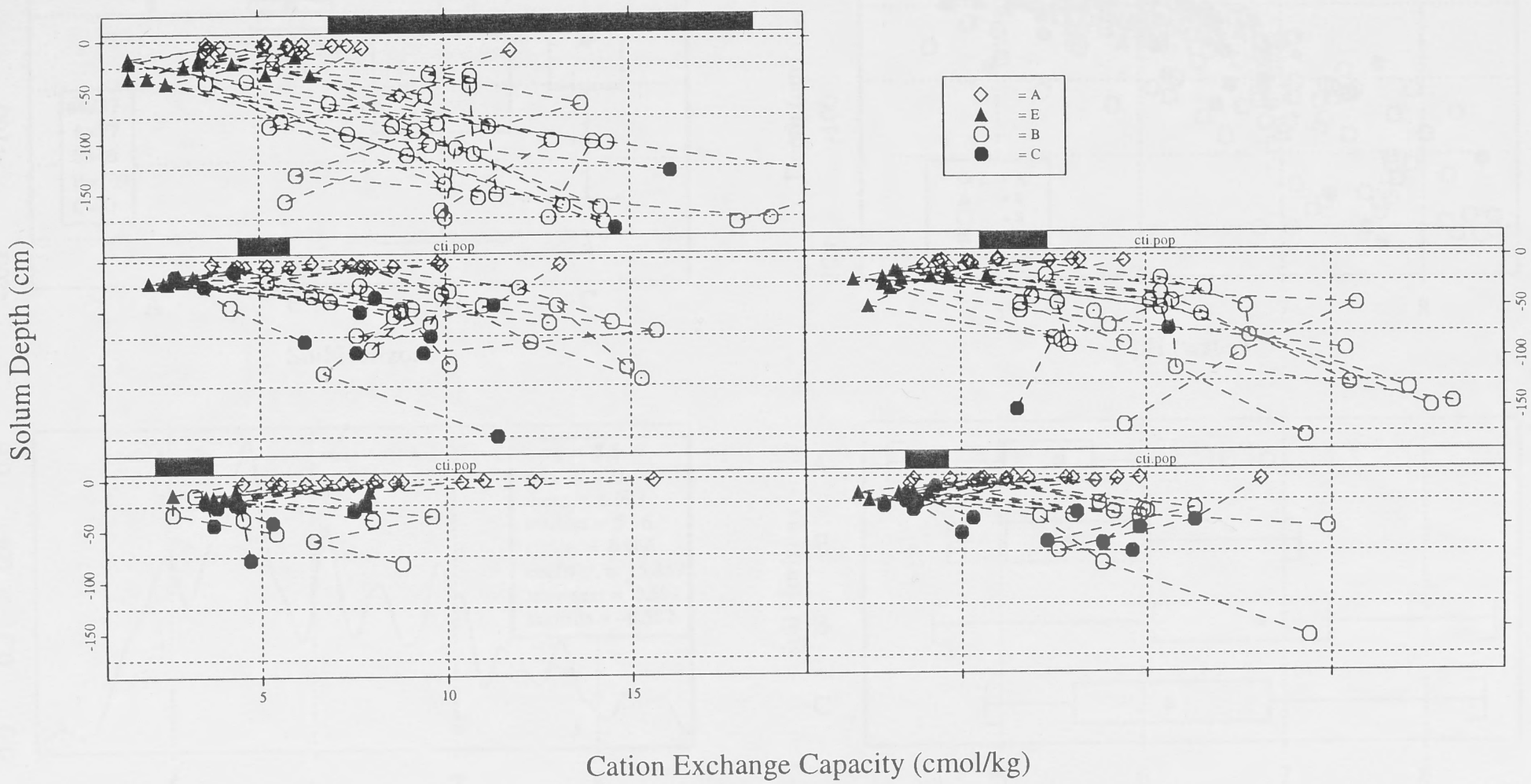
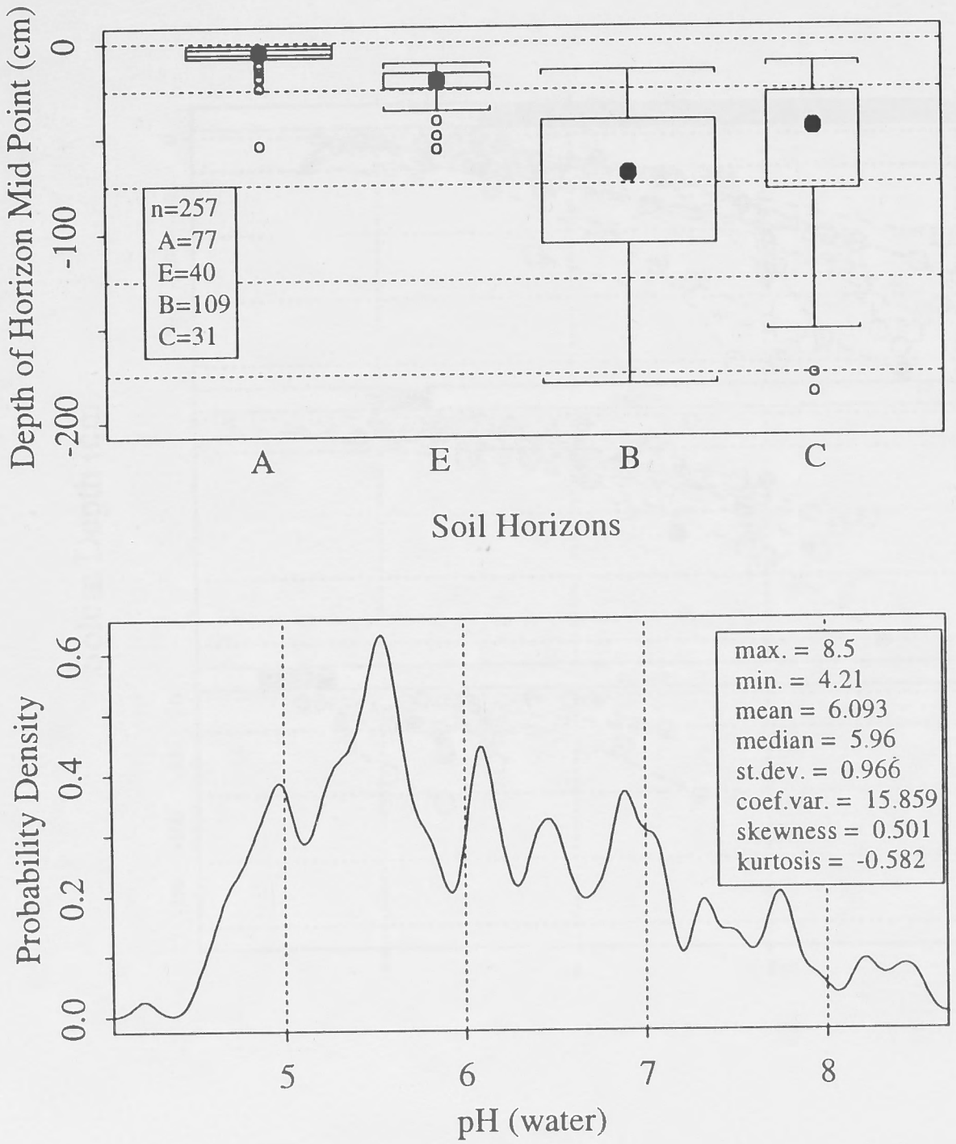


Figure A-12 Cation Exchange Capacity Trellis Conditioned by Compound Topographic Index (Ladysmith)



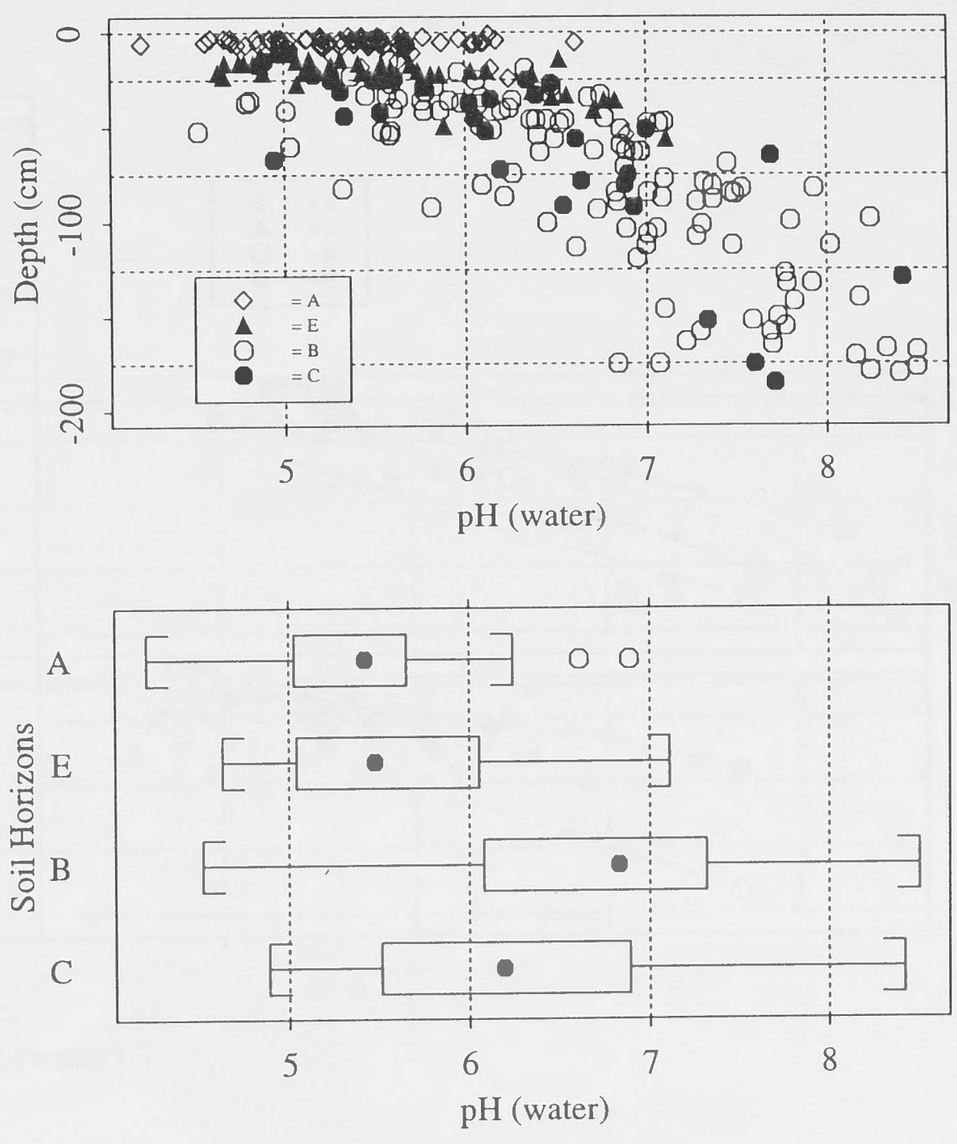


Figure A-13 pH Univariate and Bivariate EDA (Ladysmith)

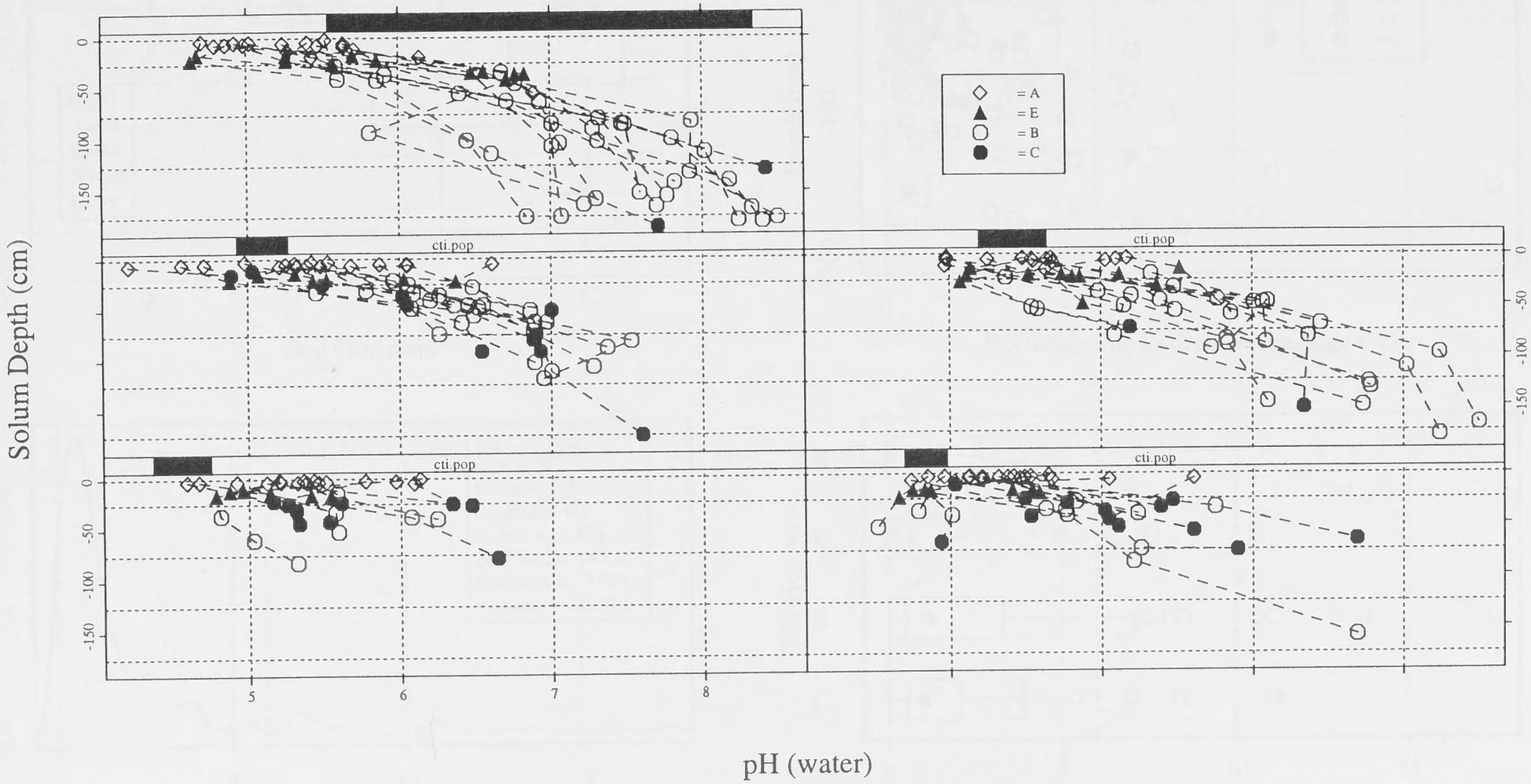
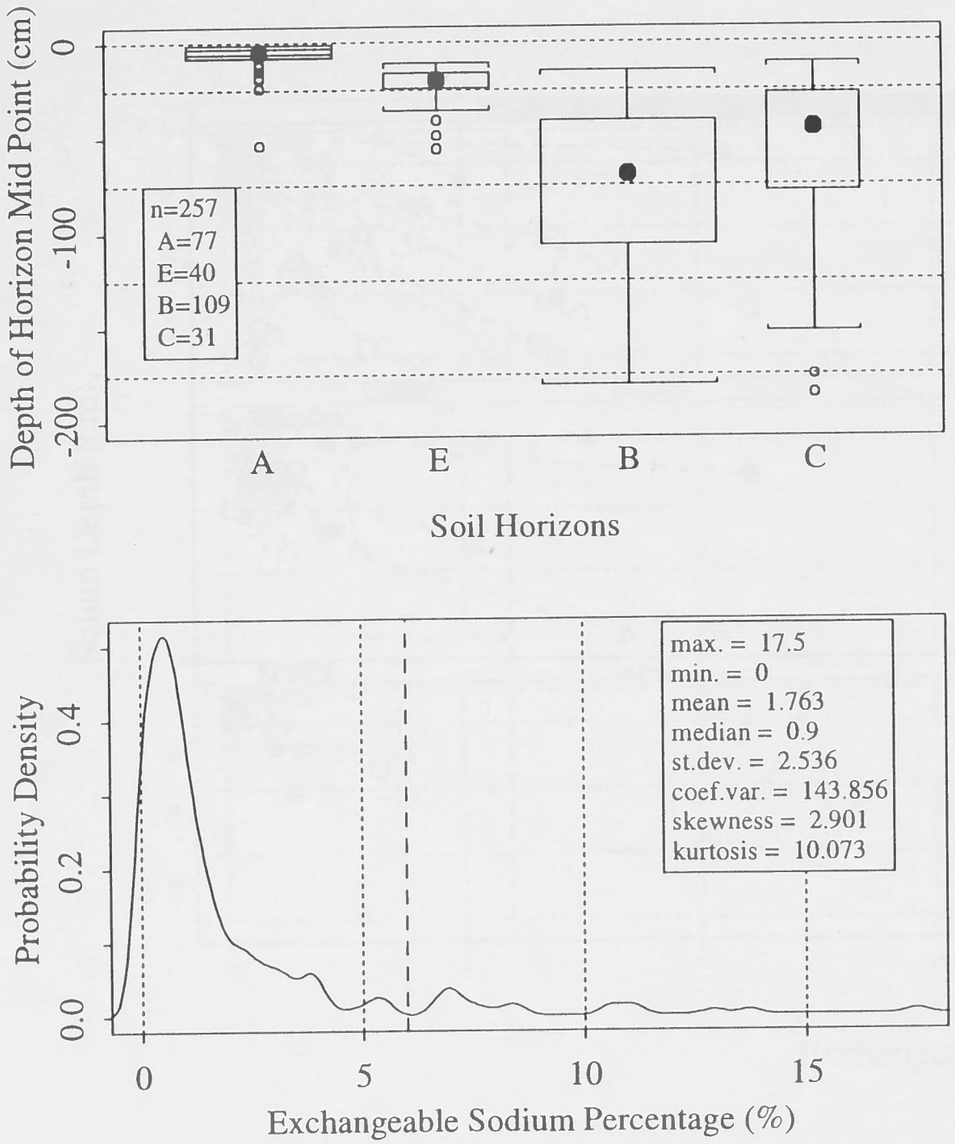


Figure A-14 pH Trellis Conditioned by Compound Topographic Index (Ladysmith)



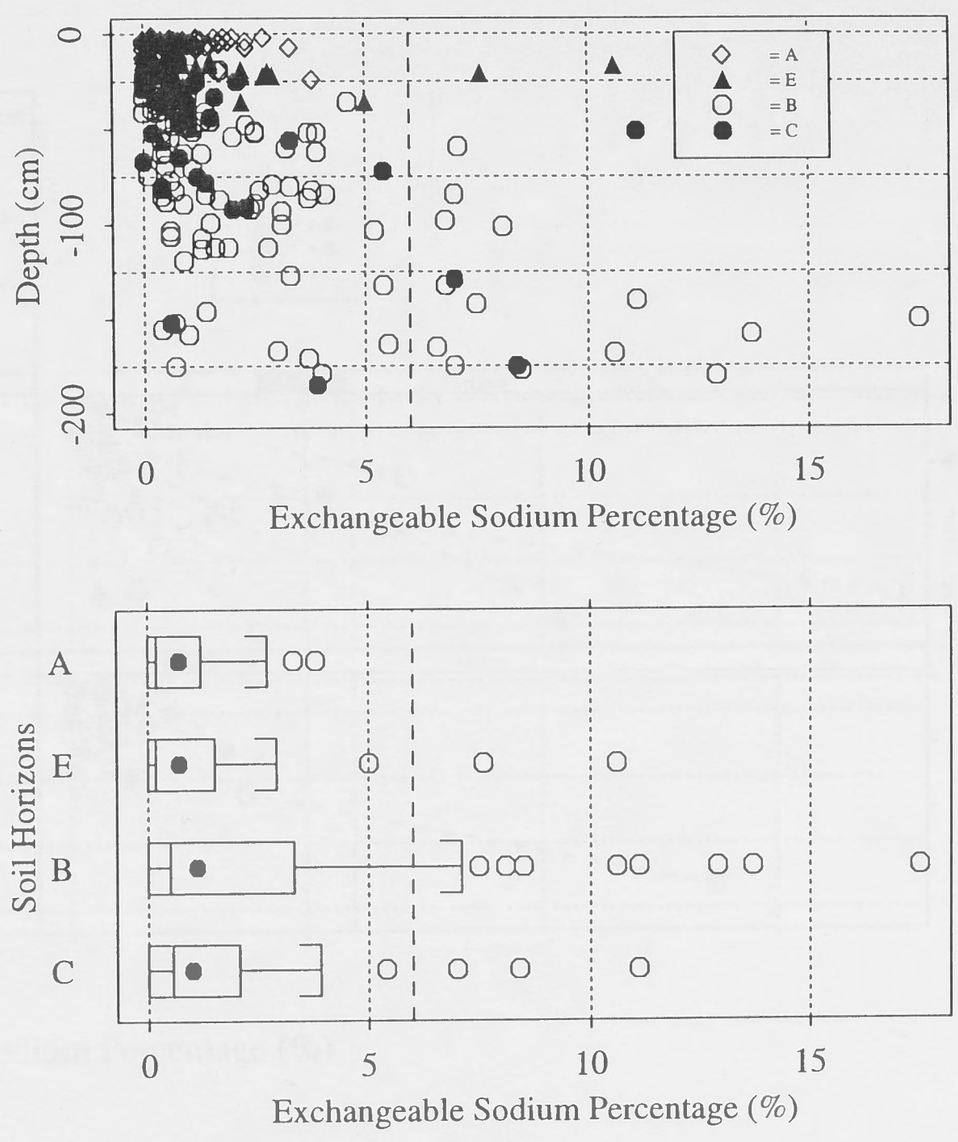


Figure A-15 Exchangeable Sodium Percentage Univariate and Bivariate EDA (Ladysmith)

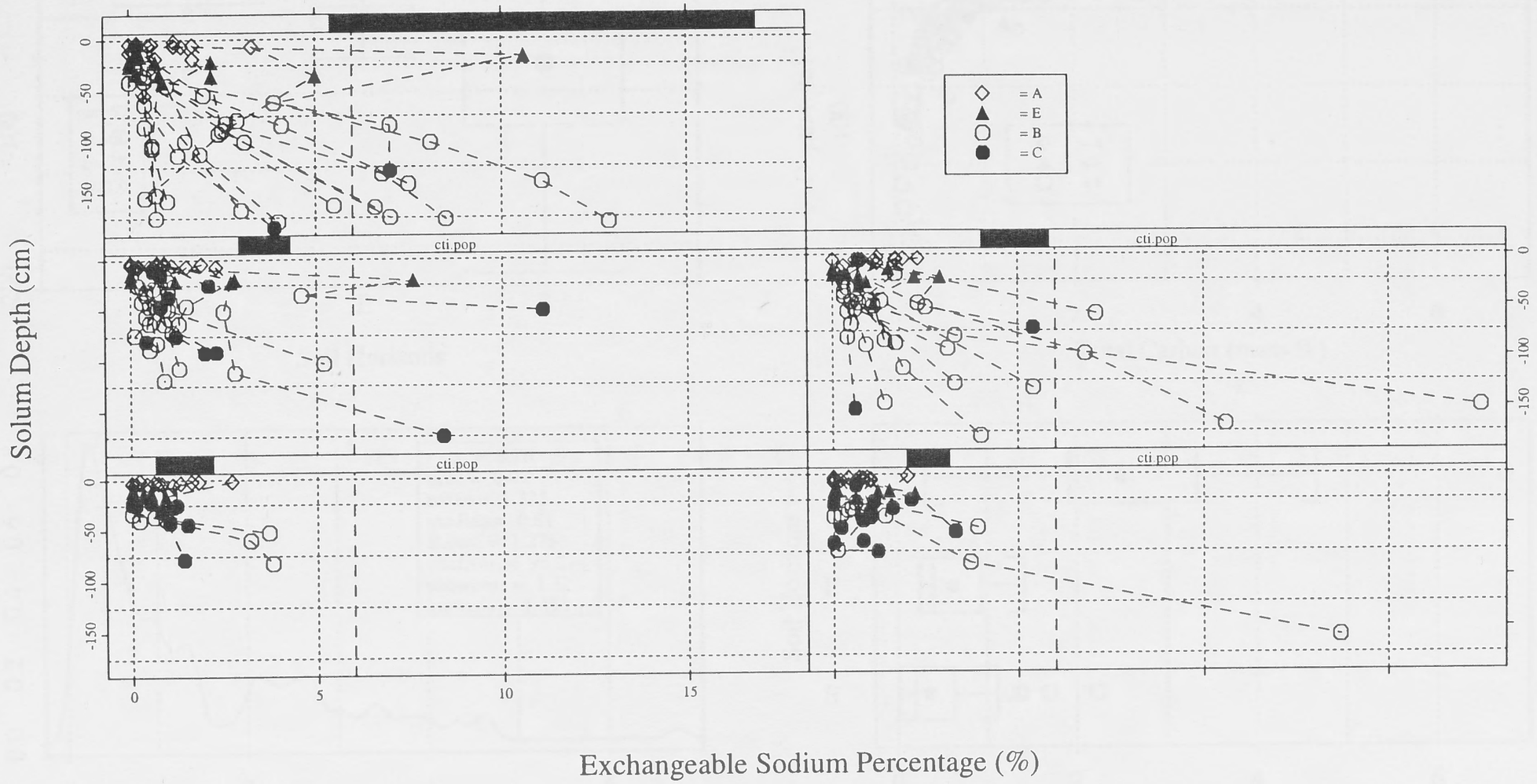
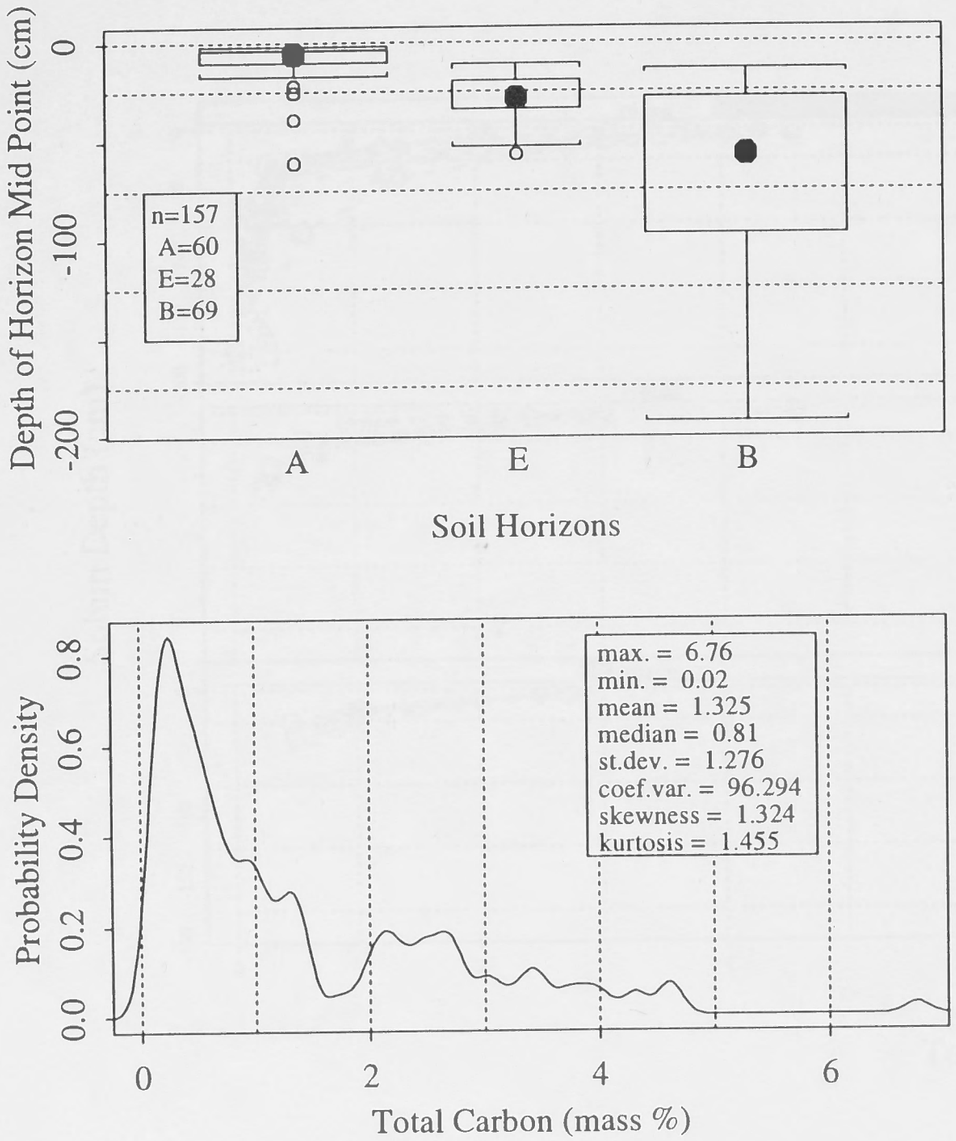


Figure A-16 Exchangeable Sodium Percentage Trellis Conditioned by Compound Topographic Index (Ladysmith)



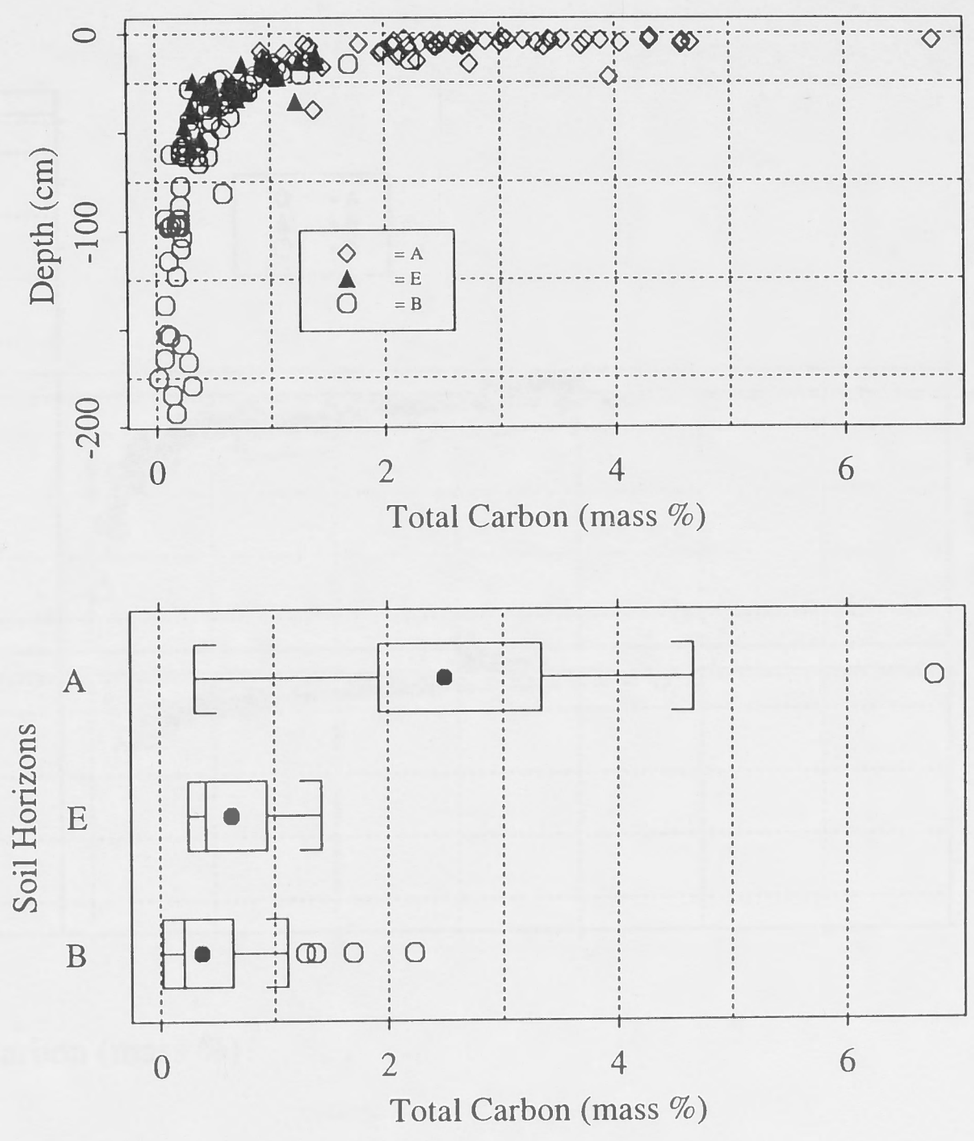


Figure A-17 Total Carbon Univariate and Bivariate EDA (Griggward)

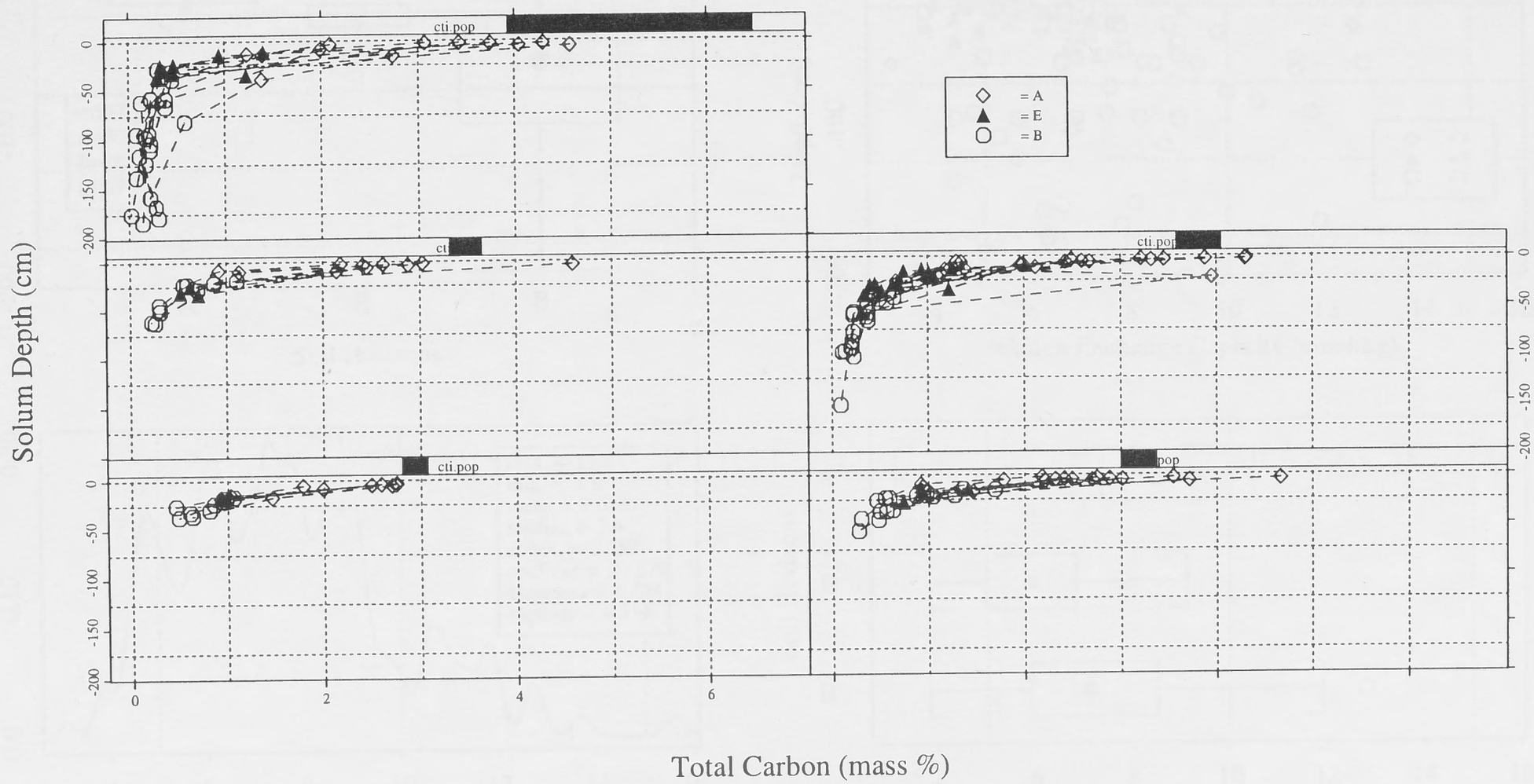
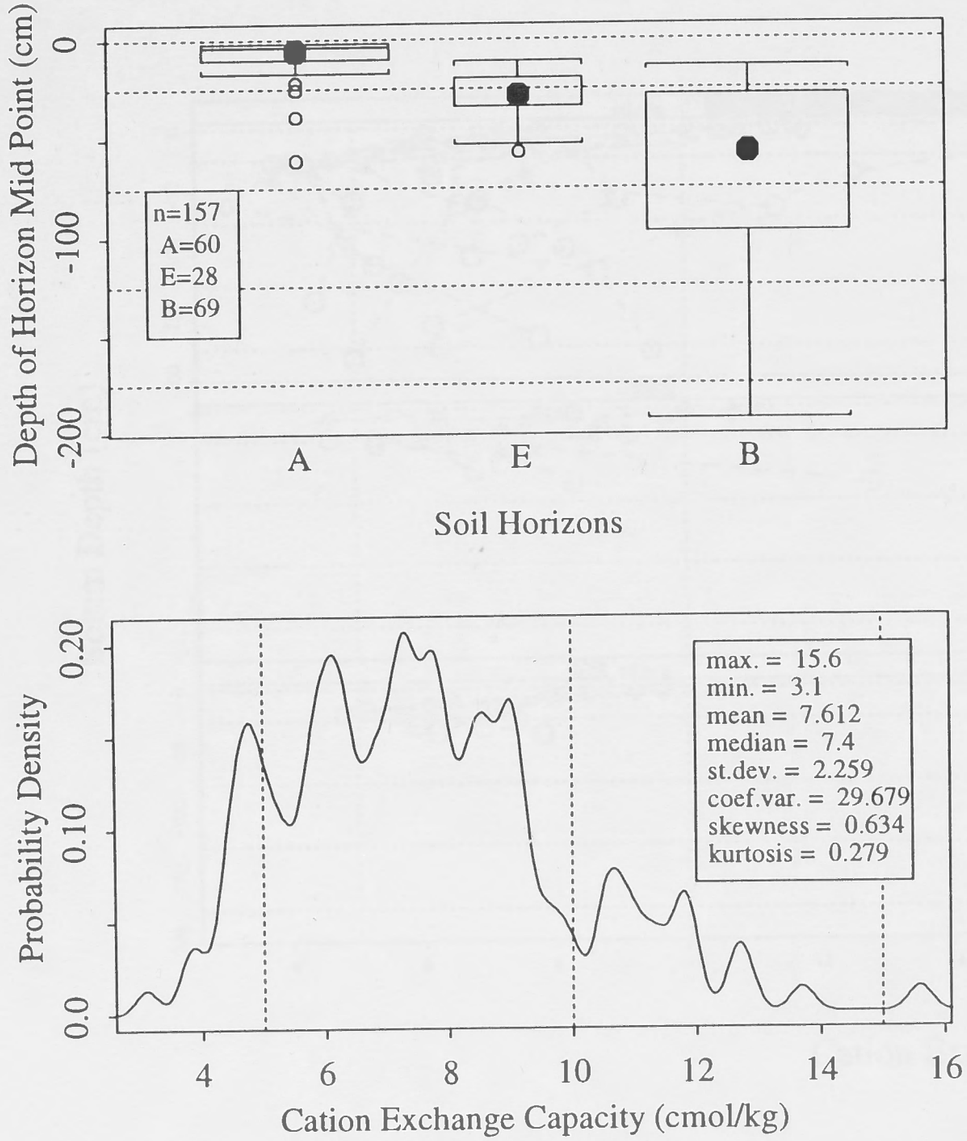


Figure A-18 Total Carbon Trellis Conditioned by Compound Topographic Index (Griggward)



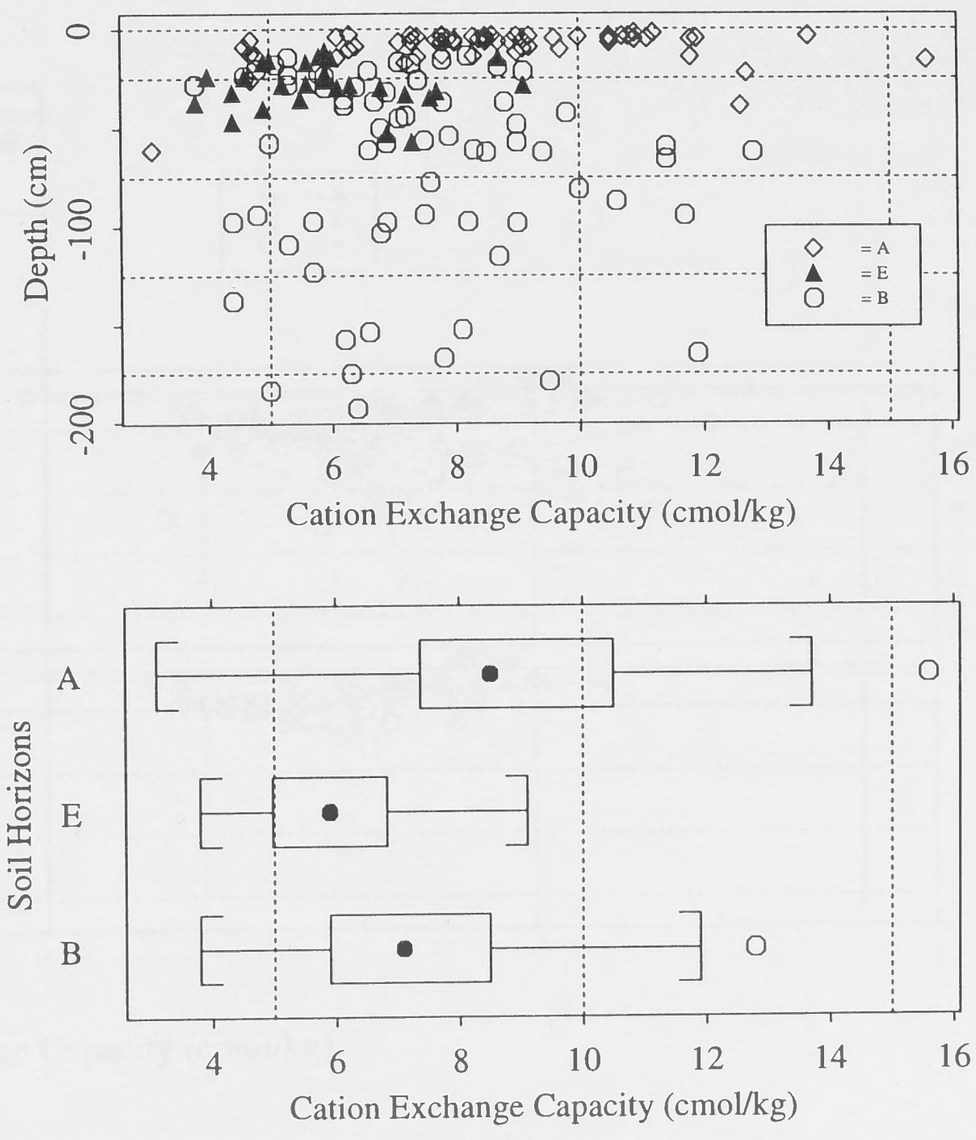


Figure A-19 Cation Exchange Capacity Univariate and Bivariate EDA (Griggward)

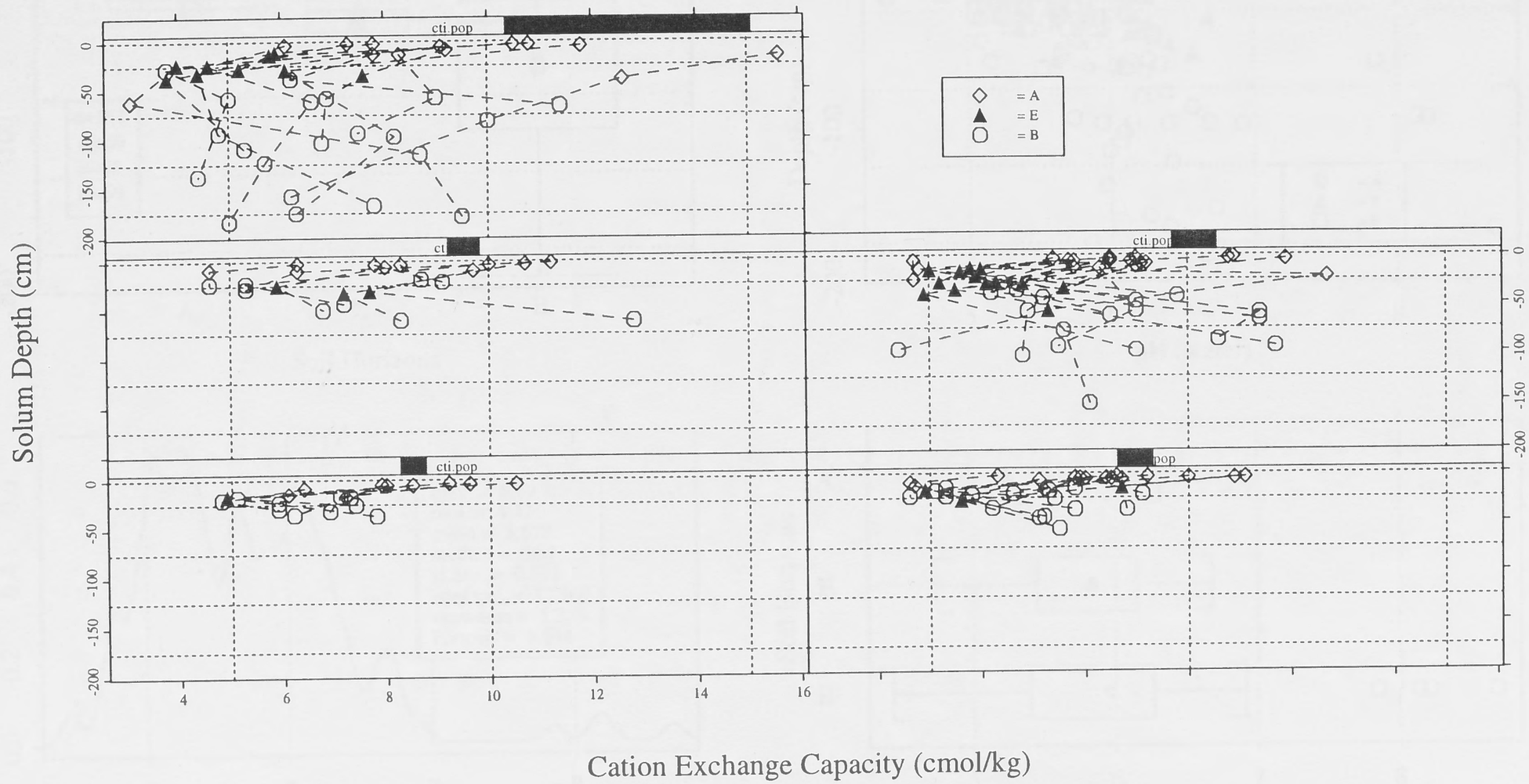


Figure A-20 Cation Exchange Capacity Trellis Conditioned by Compound Topographic Index (Griggward)

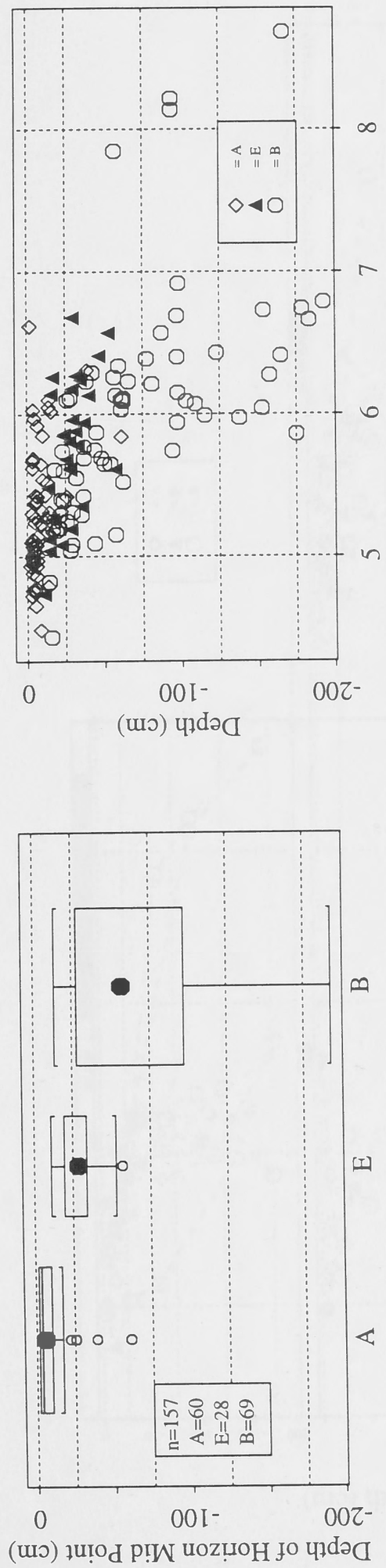


Figure A-21 pH Univariate and Bivariate EDA (Griggward)

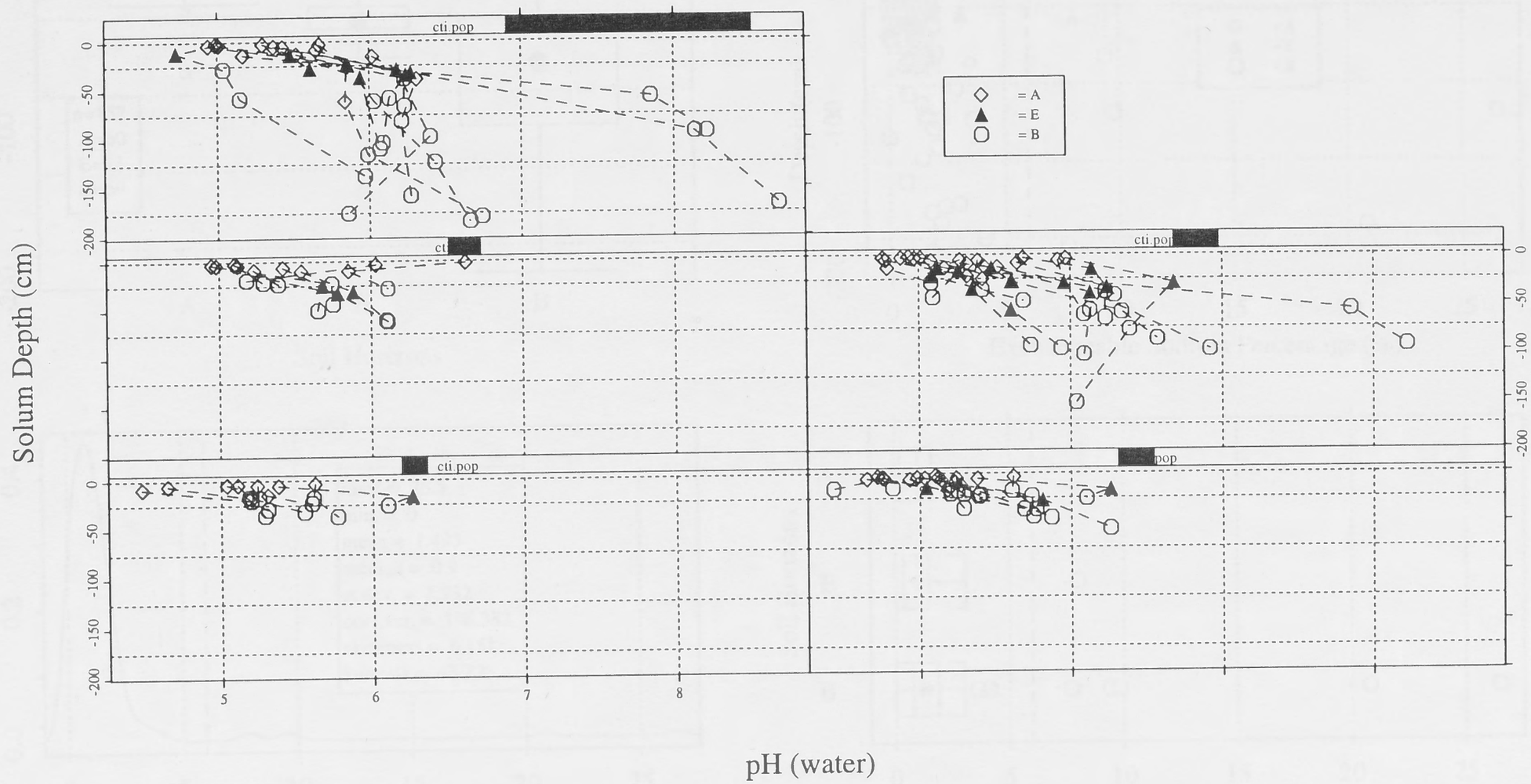
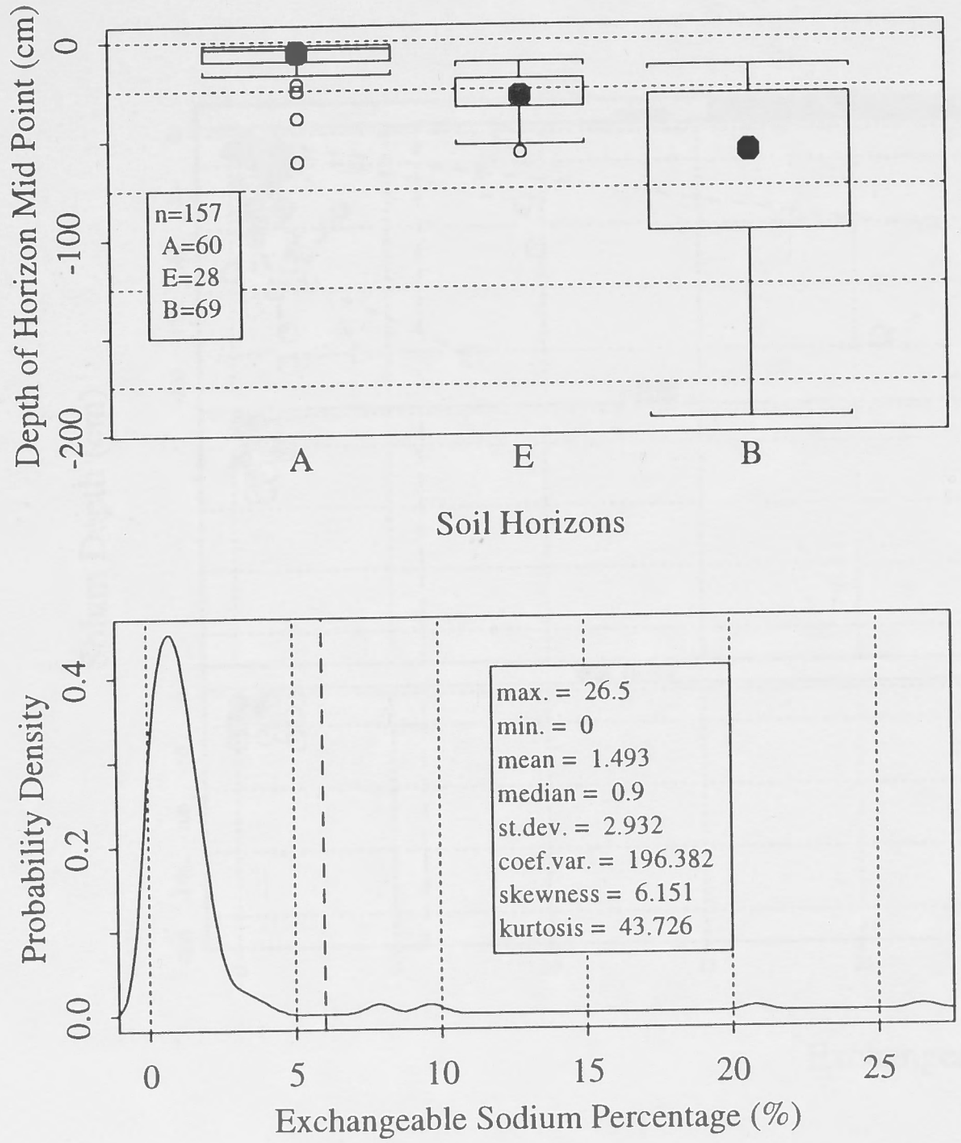


Figure A-22 pH Trellis Conditioned by Compound Topographic Index (Griggward)



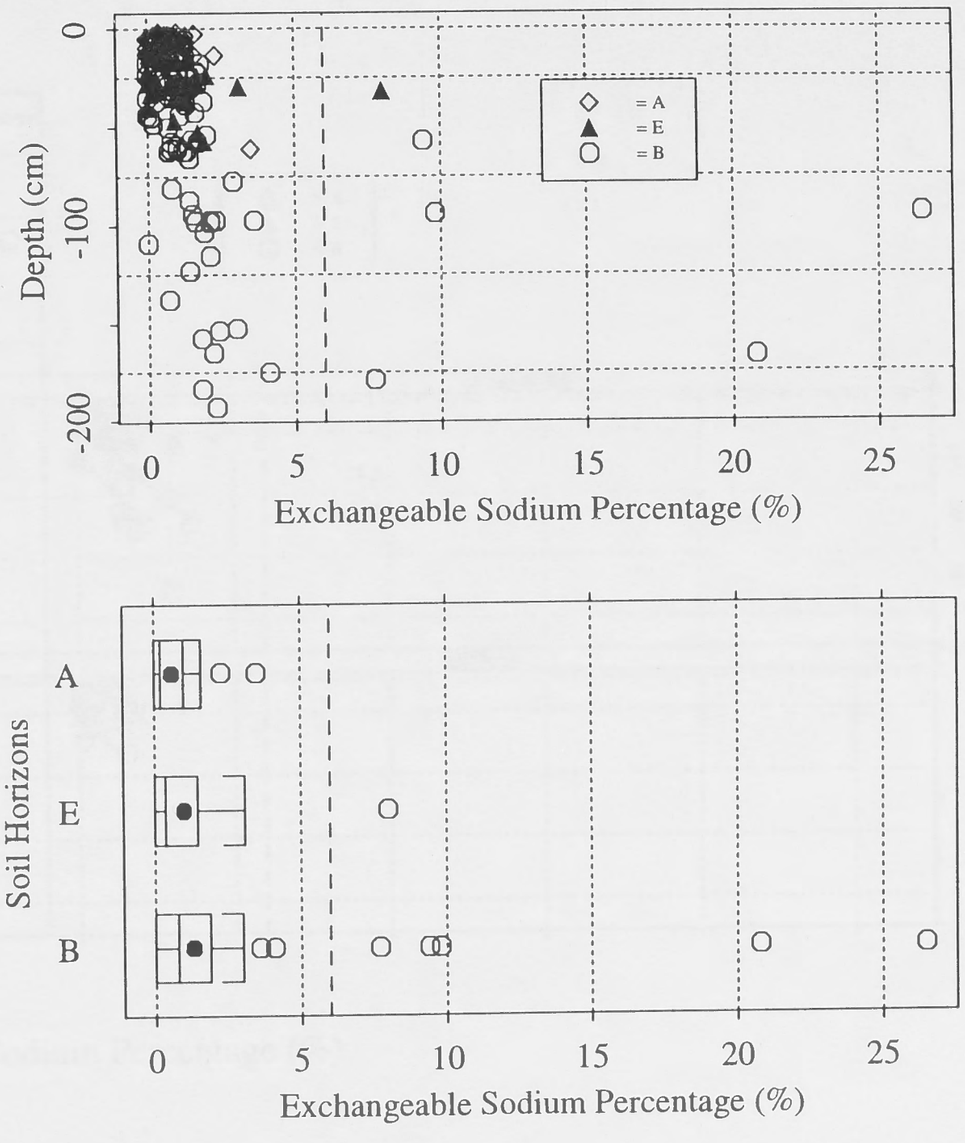


Figure A-23 Exchangeable Sodium Percentage Univariate and Bivariate EDA (Griggward)

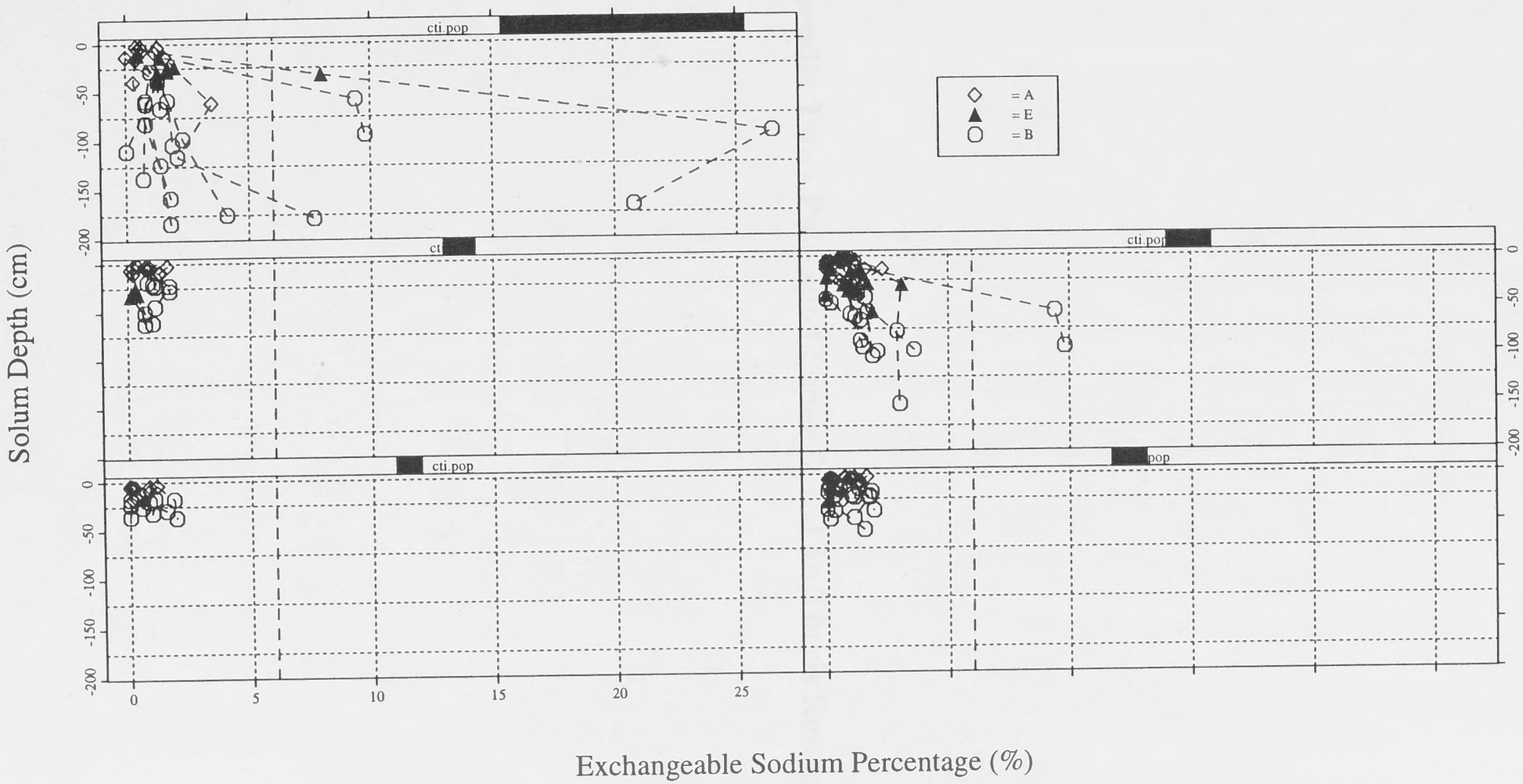


Figure A-24 Exchangeable Sodium Percentage Trellis Conditioned by Compound Topographic Index (Griggward)

Appendix Two - Published Paper Based on Thesis Research

INT. J. GEOGRAPHICAL INFORMATION SYSTEMS, 1995, VOL. 9, NO. 4, 421–432

Research Article

Soil-landscape modelling and spatial prediction of soil attributes

P. E. GESSLER
CSIRO Division of Soils, GPO Box 639, Canberra, ACT 2601, Australia and Centre for Resource & Environmental Studies, Australian National University, Canberra, Australia
email: paulg@cbr.soils.csiro.au

I. D. MOORE (deceased)
Centre for Resource & Environmental Studies, Australian National University, Canberra, Australia

N. J. McKENZIE
CSIRO Division of Soils, GPO Box 639, Canberra, ACT 2601, Australia

and P. J. RYAN
CSIRO Division of Forestry, Canberra, Australia

Abstract. Explicit and quantitative models for the spatial prediction of soil and landscape attributes are required for environmental modelling and management. In this study, advances in the spatial representation of hydrological and geomorphological processes using terrain analysis techniques are integrated with the development of a field sampling and soil-landscape model building strategy. Statistical models are developed using relationships between terrain attributes (plan curvature, compound topographic index, upslope mean plan curvature) and soil attributes (A horizon depth, Solum depth, E horizon presence/absence) in an area with uniform geology and geomorphic history. These techniques seem to provide appropriate methodologies for spatial prediction and understanding soil landscape processes.

1. Introduction

Environmental models require spatial representation of soils because they modify material and energy fluxes at the earth's surface. Ideally, spatial predictions of soil layers, individual soil attributes and, eventually, soil-landscape processes, are needed at a scale appropriate for environmental management (Moore *et al.* 1993). The challenge is to develop explicit, quantitative, and spatially realistic models of the soil-landscape continuum useful for a variety of purposes beyond taxonomic classification (McSweeney *et al.* 1994). A promising development is the potential for correlating soil attributes with terrain and environmental attributes that are simple to measure and have physical meaning (Moore *et al.* 1993, McKenzie and Austin 1993). The underlying hypothesis of Moore *et al.* (1993) was that the development of soil toposequences often occurs in response to the way water moves through and over, the landscape. Water movement is in turn controlled by the geometry of the land surface and underlying materials. The geometry of the land surface, therefore, can be used as a first approximation for predicting the movement of water and related material (Moore *et al.* 1991).

0269–3798/95 $10·00

There has been a trend in recent work (Bouma 1989, Gessler *et al.* 1989, Baize and Girard 1992, FitzPatrick 1993) towards using soil layers rather than soil profiles or pedons as the basic object for study. Soil layers may have a pedogenic (soil horizon) or geomorphic (stratigraphic unit) origin. Regardless of origin, they form a logical building block for spatial modelling and interpretation of how sequences of layers behave. The soil layers at any location are a result of integrated pedo-geomorphic and hydrological processes (Simonson 1959, Butler 1964). As such, a description of the arrangement, dimension and nature of the soil layers at locations in the landscape may be used as a link or pointer to the spatial distribution of processes and vice-versa.

However, soil-landscape processes operate across a range of spatial and temporal scales (Allen and Starr 1982, Kachanowski 1988) and it is clear that imprinting of past climates, truncation by over-riding processes, and process synergisms occur (Malanson *et al.* 1990, Allison 1991). Consequently, soil attributes exhibit different and complex scales of variation (Butler 1964, Beckett and Webster 1971, Burrough 1993). Thus, our expectations for deciphering the relationship between pattern and process should vary within and between physiographic domains. This reinforces the need to develop environmental correlations using exploratory data analysis (Tukey 1977) followed by explicit definition based on physically interpretable statistical models (Chambers and Hastie 1992, McKenzie and Austin 1993).

The general form of such statistical models being:

$$S_i = f_i(\text{slope, catchment position, solar radiation, gamma radiometrics, } \ldots)$$

where:

- S is individual soil attribute (e.g. soil depth, pH, etc.);
- f is a function of one or several environmental attributes;
- i is the physiographic domain characterized by common climate, parent material, geomorphic history, vegetation, etc.

In this approach, the statistical model is developed using data from measurements of soil attributes (response variable) made in the field at locations where measurements of environmental attributes (explanatory variable(s)) are available. Spatial prediction is then achieved using environmental variables, such as slope, that may be generated using digital terrain methods or other techniques. The environmental variables must be easier to obtain than soil variables and be available for the complete study area, otherwise, intensive sampling of soil variables in association with an interpolation or surface fitting procedure would be a more efficient method for spatial prediction. The definition of the physiographic domain where a developed model applies depends on the scale and purpose of the work (McSweeney *et al.* 1994). It could be for broadly defined regions (e.g. river basins, land systems) or more local areas defined by hillslopes within a given geomorphic unit. With this approach, primary data can be re-analysed with different combinations of response and explanatory variables, and statistical methods can be varied as suggested by exploratory data analysis and general field observation.

Australia contains vast areas with scant land resource information. The resources for collecting basic data sets to understand environmental function and management are limited (McKenzie 1991). This paper presents initial results on the testing of a method for developing explicit soil-landscape models using pedological knowledge, spatial analysis, field sampling, exploratory data analysis and statistical modelling techniques. The broad aims of this work were to develop:

(*a*) Procedures for the quantitative characterization of landform because of its importance as a local scale predictor of soil attributes.
(*b*) A rational and efficient soil sampling strategy.
(*c*) Robust statistical models for the spatial prediction of soil attributes in an area with uniform geology and geomorphic history; and
(*d*) Quantitative methods for comparing and understanding soil-landscape processes.

Developed models and predictions may then be used to parameterize other models for environmental management (e.g. estimation of erosion hazard, crop growth, water quality, nutrient cycling) and for simulation of impacts due to changes in land use.

2. 2-D spatial characterization of processes: terrain analysis

Moore *et al.* (1991) review terrain analysis and its application in the earth sciences. Primary and secondary (or compound) topographic attributes are recognized and they present a table summarizing the significance of these attributes for characterizing the spatial distribution of landscape processes. Many of the attributes have potential use as spatial predictors of soil attributes. Primary attributes are directly calculated from elevation data and include areal measures such as specific catchment area and point measures including the first and second derivatives such as slope, aspect, plan and profile curvature. Secondary attributes involve combinations of the primary attributes that quantify the contextual nature of points or characterize the spatial variability of specific processes occurring in the landscape or both. Methods of computation are presented by Moore *et al.* (1991, 1993).

Digital topographic attributes are scale dependent and if these effects are not considered, computed attributes may be meaningless or the processes of interest may be masked (Moore *et al.* 1991, 1994). Moore *et al.* (1994) report critical differences in the computation methods of primary and secondary topographic attributes and, for example, advise against the use of the D8 method of flow direction computation. This method does not allow flow dispersion and produces unrealistic flow patterns. This significantly influences the computation of flow accumulation which is critical to the computation of many spatial hydrological and soil-landscape attributes such as catchment and dispersal areas. Differences in environmental attribute correlations and model development will occur due to physiographic setting, scale of analysis, computation methods, and others (data structure, quality and error). It is essential for a modelling framework to have explicit definition of decisions relating to the particular combination of methods applied (McSweeney *et al.* 1994, Wagenet *et al.* 1994) so that other workers can evaluate, repeat or improve the model.

3. Material and methods

3.1. *Study region*

The study region is the Wagga Wagga 1:100 000 topographic map sheet located on the western slopes of the Great Dividing Range in southeastern Australia (147°E, 35°S; 147°E, 35°30′S; 147°30′E, 35°30′S; 147°30′E, 35°S). This region was chosen because it has a diverse range of geological units, landforms, soils and land uses typical of the broader Murray-Darling River Basin. Areas with distinct combinations of geology and landform (physiographic domains) have been delineated and later work will develop soil-landscape models in each area for testing and comparison. This paper focuses more specifically on initial methodology development in a 100 km^2 pilot study

area (centered on 147°27′E, 35°24′S) dominated by gently-rolling erosional landforms on Ordovician metasediments. The dominant land use is pastoral grazing.

3.2. *Soil-landscape model development*

Two methods have recently been proposed for development of explicit and quantitative soil-landscape models (McKenzie and Austin 1993; McSweeney *et al.* 1994). Both methods are similar in approach and require a definition of purpose, scale of application and stratification of the physiographic domain for field sampling. The work reported here is aimed at developing spatial models of soil layer patterns within the Ordovician meta-sediment physiographic domain. The models of soil layer patterns are viewed as critical if they are to lead to eventual spatial prediction of individual soil attributes and soil-landscape processes in three and four (time) dimensions. The scale of application is the hillslope within small catchments and is intended to provide information at the local land management level in the study area. The soil layer is used as the basic object of study and the catchment is the boundary of the system, due to its significance for spatially related hydrological and erosional processes.

Stratification into distinct physiographic domains for soil-landscape model development is a critical initial step. The quality of stratification depends on the availability of prior information on soils, geology, vegetation, landform and surficial materials. At the onset of this work a 1 : 100 000 geology map (Raymond 1992) was generated and initial stratification into physiographic domains was performed using these data. Additional data layers (soils, landform, stratigraphy, vegetation, climate) are being generated as part of a collaborative project, and subsequent work will look at stratification using these integrated data more specifically. The focus here is on the methods of explicit soil-landscape model development within one physiographic domain.

Digital contours (10 m contour interval), streamlines and spot heights registered to the Australian Map Grid (AMG-UTM) were obtained from the New South Wales Land Information Centre in digital form. A base-line 20 m × 20 m grid digital elevation model (DEM) for the 100 km^2 study area was generated using the program ANUDEM (Hutchinson 1989). Scaling parameters, fractal and error properties of this surface are reported elsewhere (Moore *et al.* 1994). Seventeen catchments were delineated and a full range of primary and secondary topographic attributes were generated for each catchment using the methods of Moore *et al.* (1993, 1994). Flow or area accumulation (i.e. specific catchment area) was calculated using the FRho8 flow dispersion algorithm (Moore *et al.* 1994), and a 100 cell channel initiation threshold. When this threshold is reached, flow accumulation switches from dispersive to channelized flow using the D8 method.

As part of this work, a new algorithm was developed for computing upslope statistical moments (i.e., upslope mean slope, upslope mean plan curvature etc.,) to provide quantitative information about the upslope catchment area feeding into each grid cell. The S-Plus language (Statistical Sciences 1992) was used to develop a function to create graphical displays of the probability density function and listing of the statistical moments and Moran statistics (Goodchild 1986) for each terrain attribute on a catchment basis. This function enables the rapid characterization of a catchment and quantitative comparison of the overall differences between catchments or specific zones within a catchment. A contiguous five catchment subarea (20 868 cells or 834·72 hectares) was selected for soil-landscape model development because it encompassed

a range of the topographic variability (including aspect) characteristic of the physiographic domain as a whole.

3.3. *Development of an explicit and quantitative sampling strategy*

An iterative sampling strategy using four criteria was used to select field sample sites. First, the sampling plan aimed to reflect the provisional predictive pedologic model by sampling evenly along the predictive variable(s) in attribute space. Secondly, randomization was used to achieve an unbiased sample. Thirdly, sampling inefficiencies due to spatial dependence in soil attributes were minimized. Fourthly, locational error between the digital terrain model and the real world was minimized.

3.3.1. *Provisional predictive pedologic model & randomization*

When a soil surveyor initiates a survey in un-mapped territory, he or she often begins with an implied model or mental construct and begins testing hypotheses with sample points. This provisional predictive pedologic model (McKenzie and Austin 1993) evolves as points are sampled. But much of this information about continuous soil-landscape variation is lost or subsumed when map unit lines are drawn. A common provisional model is the catena (Latin = a chain) soil-landscape (Milne 1935) that implies a concordance of soil pattern with landform as one traverses from hilltop to valley bottom along toposequences. The compound topographic index (CTI), often referred to as the steady-state wetness index, is a quantification of catenary landscape position. It is defined as:

$$\mathrm{CTI} = \ln\,(A_s/\tan\beta) \tag{1}$$

where A_s is the specific catchment area (area (m^2) per unit width orthogonal to the flow direction) and β is the slope angle. Moore *et al.* (1993) showed that the CTI is correlated with several soil attributes such as silt percentage ($r = 0{\cdot}61$), organic matter content ($r = 0{\cdot}57$), phosphorus ($r = 0{\cdot}53$) and A horizon depth ($r = 0{\cdot}55$) in the soil surface of a small toposequence. The CTI was used in this work as an explicit and quantitative provisional predictive pedologic model. To develop a robust statistical model for testing hypothesized correlations, it is sensible to sample evenly in CTI attribute space. Thus, the CTI was divided evenly into five equal percentile classes (figure 1 (*a*)). The goal of this work was to develop a soil-landscape model applicable to the broader Ordovician metasediment physiographic domain. Therefore, the percentile break-points were computed using all the grid cells falling on this bedrock type in the 100 km^2 study area. Figure 1 (*b*) shows a spatial display of the percentile classes for the study catchments. The percentile classes also provide convenient strata or patches that can be used for randomization to meet the second sampling criterion.

3.3.2. *Spatial dependence*

Soil attributes show varying degrees of spatial dependence and this reduces the efficiency of random sampling (McBratney and Webster 1983). Spacing sample sites using information about the spatial dependence structure increases the information content of samples. No *a priori* information on the spatial dependence structure of the soil attributes of interest was available. Instead we postulated that the spatial dependence structure of the CTI related in a general way to the spatial dependence structure of the soil attributes of interest. Moran's I coefficient (Goodchild 1986), which characterizes the overall strength of spatial dependence, is 0·70 for the CTI cells on the Ordovician meta-sediments in the 100 km^2 study area. This indicates strong

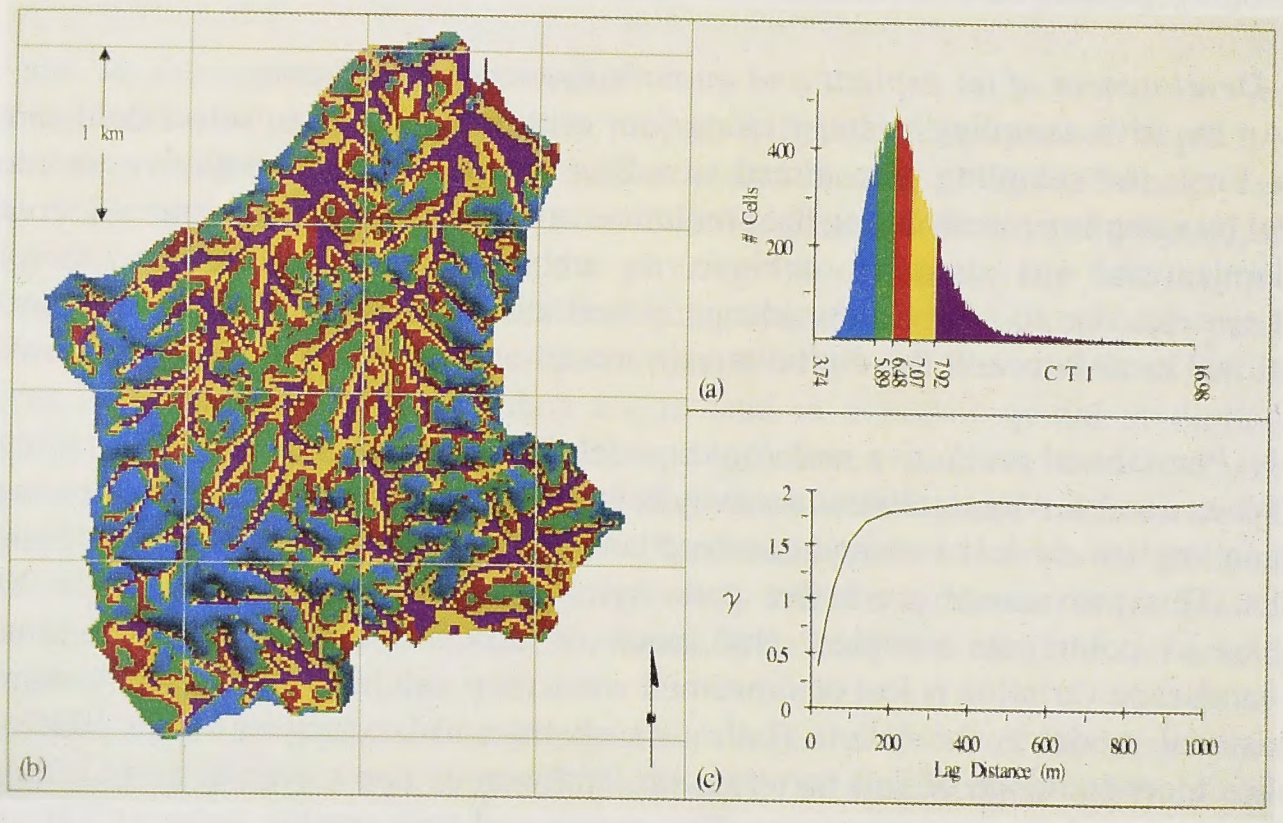


Figure 1. (*a*) Twenty percentile histogram of CTI, (*b*) spatial display of CTI for study catchments and (*c*) CTI variogram.

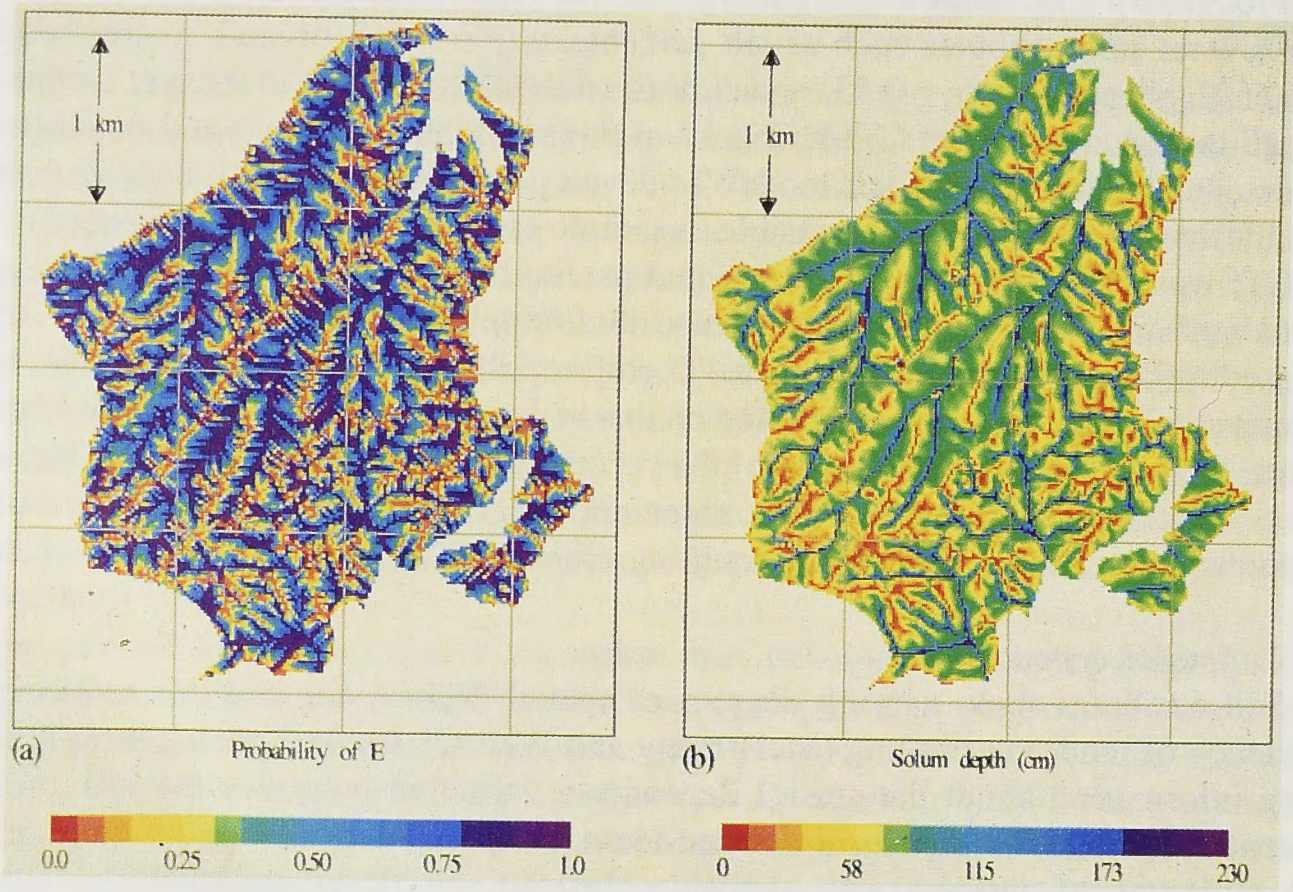


Figure 3. (*a*) Probability of E horizon and (*b*) predicted solum depth.

spatial dependence in the CTI. The variogram (Webster and Oliver 1990) is a common method of quantifying the spatial dependence structure of a regionalized variable (Matheron 1971). Figure 1 (*c*) shows the computed variogram for the CTI cells on the Ordovician metasediments. This variogram shows a range (distance within which spatial dependence occurs) of approximately 500 m. This suggests that statistical independence can best be maintained by spacing samples 500 m or more apart. An assumption is that the spatial dependence is stationary across the landscape. Subsequent sampling may be useful at nested scales within this distance to develop a useful understanding of short-range variation for individual soil attributes. Short range variation was not of primary interest and will not be discussed in this paper.

3.3.3. *Location of sample sites*

Accurate location of field sample points allocated using a geographical information system (GIS) is critical to the development of robust statistical models. To minimize locational errors, samples were located only in attribute patches with a minimum size of 3×3 grid cells (0·36 ha). This was accomplished by using a two cell erosion and dilation procedure to eliminate thin areas and small patches.

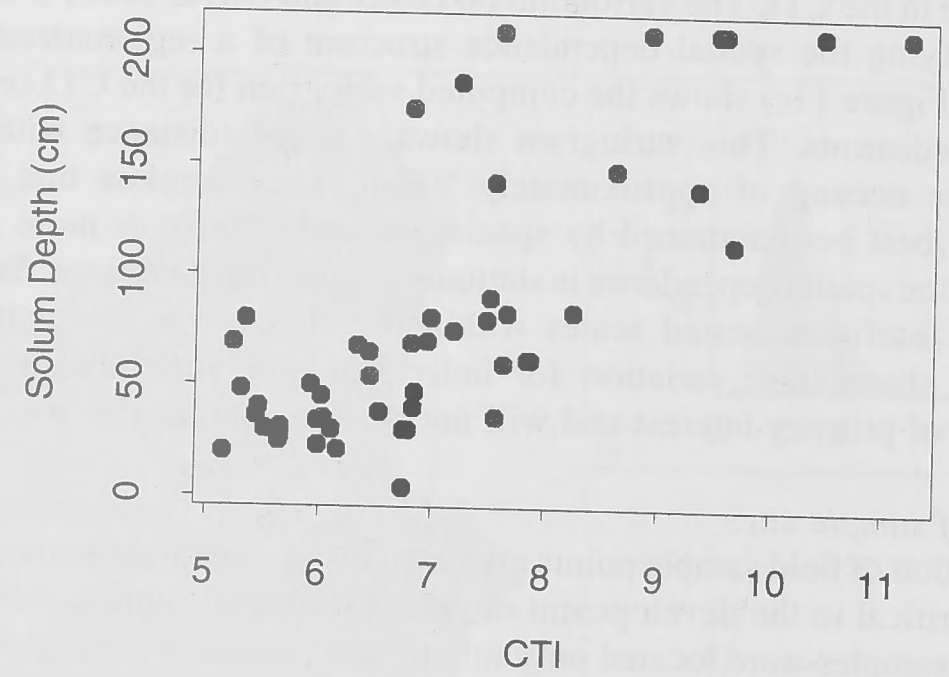
3.4. *Sample site allocation and data collection*

Sites were allocated in two batches of 30 samples. Six samples were distributed in each CTI percentile class according to the following iterative scheme. The patches for each class were numbered from 1 to n (total number of patches for percentile class). A random number generator was used to produce a random number vector of length n. Sites were selected sequentially from randomly selected patches and Australian Map Grid coordinates produced for each site. Sites within 500 m of previously selected sites were discarded and the next random patch selected until six sites were allocated for each class. Each site was located in the field using a global positioning satellite (GPS) receiver. The slope, aspect, elevation and specific catchment area attributes for each site were output from the GIS and used in the field to refine site placement and ensure consistency. At each site a 71 mm diameter core was taken to a maximum depth of 2·3 m. The cores were described according to McDonald *et al.* (1990).

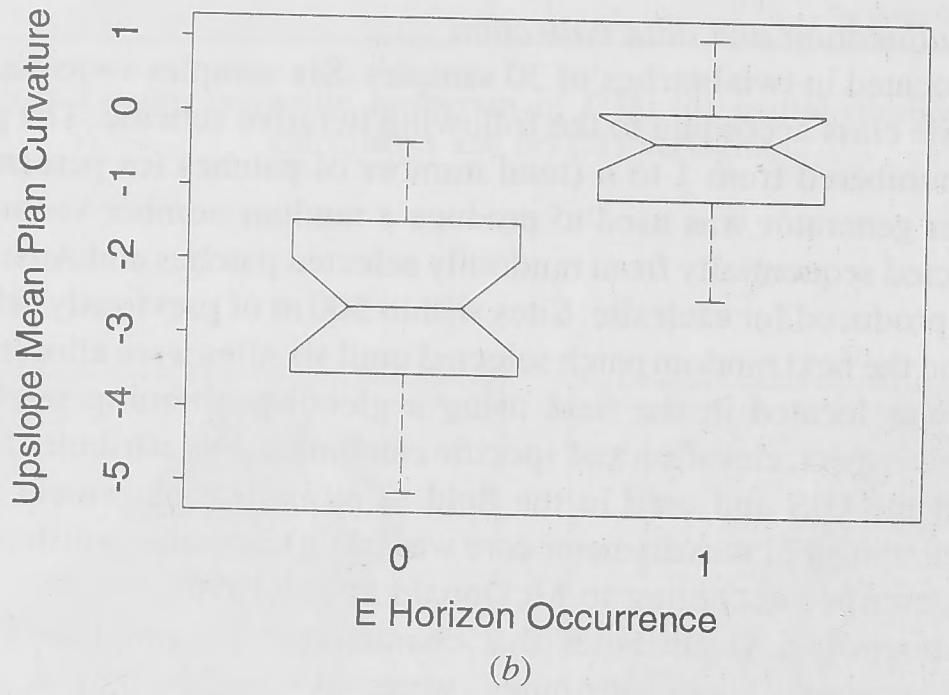
Diagnostic morphological attributes that characterize the soil layers were used for model development. These attributes were: A horizon depth, E horizon presence/absence, E horizon depth, mottle presence/absence, depth to mottles, A horizon clay percentage, B horizon clay percentage and solum depth (A + E + B horizon depths). Results pertaining only to A horizon depth, solum depth and the probability of encountering an E horizon are presented to demonstrate the methodology. The A horizon depth is a general guide to nutrient status of soils in the study area and also an indicator of surface stability to erosional and depositional processes. An E horizon is indicative of downward or lateral percolation and leaching processes and periods of water logging. This has an impact on biological productivity and trafficability. Solum depth provides an indication of the available water capacity, and also exerts a major control on biological productivity.

3.5. *Exploratory data analysis and statistical model development*

A matrix of scatter plots (Cleveland 1993) was developed to identify patterns or structures within the data and to provide an indication of soil and terrain attribute correlations. Figure 2 (*a*) shows a scatter plot of solum depth versus CTI and figure 2 (*b*) a box plot of upslope mean plan curvature versus E horizon presence (1) or



(*a*)



(*b*)

Figure 2. (*a*) scatterplot of solum depth versus CTI, (*b*) boxplot of upslope mean plan curvature versus absence (0) or presence (1) of an E horizon.

absence (0). This illustrates a simple visualization of relationships between soil (response) and terrain (explanatory) attributes that provided the first indication of predictive potential. This was followed by a stepwise exhaustive search technique (Statistical Sciences 1992) that considers possible subsets of explanatory variables based on the residual sum of squares. Statistical modelling was then performed using generalized linear models (McCullagh and Nelder 1989). Diagnostic methods of identifying outliers, influential observations and violations of model assumptions were used routinely (Cooke and Weisberg 1982).

Two types of generalized linear model were used. The first was a multiple regression with an identity link function and poisson error function. It is similar to a classical least squares multiple regression, except the poisson error function is specified, in this instance, because the variance increases with the fitted mean. The second type of model was used for predicting a binary response variable, in this instance, the probability of

encountering an E or bleached horizon in the upper part of the soil profile. This generalized linear model uses a logistic link function and binomial errors and is often referred to as a logistic regression model. The proportion of variation accounted for by a logistic model cannot be expressed using a statistic analogous to R^2. Model adequacy is assessed in terms of the prediction errors and the reduction in residual deviance which is distributed approximately like χ^2 (McCullagh and Nelder 1989).

4. Results and discussion

The ubiquitous and substantial short range variation of soil attributes places a fundamental limit on the quality of spatial prediction. This issue has been avoided in traditional soil surveys by the delineation of somewhat qualitative and subjective map unit lines based on morphological soil types (McSweeney *et al.* 1994). Webster (1977) concluded that the variation accounted for by a typical general purpose soil survey would range from about half the total variance for soil physical attributes to less than one tenth for some soil chemical attributes. This provides an informal measure for judging the success of a statistical model. Statistical models that predict soil attributes using topographic attributes are presented in table 1. The percent reduction in deviance provides an indication of the proportion of response variability explained by the fitted model and is similar to the R^2 for multiple regression. The results in table 1 are encouraging because of the large reductions in deviance accounted for by the fitted models.

As expected, CTI was a useful predictor because it combines contextual and site information via the upslope catchment area and slope, respectively. Plan curvature was not expected to have a strong predictive power because it does not include contextual information. However, it was significant in predicting the A horizon and solum depth in combination with CTI. This suggests that local scale pedogenic as well as hillslope scale processes are influencing soil profile development. Upslope mean plan curvature

Table 1. Regression equations for prediction of soil attributes (Standard errors are shown in parentheses)*.

Regression models				Reduction in deviance (%)
A horizon depth =	0·92 + 5·67 plancrv + 4·88 CTI			63%
SE	(14·1) (1·4)	(1·9)		
Solum depth =	− 57·95 + 12·83 plancrv + 21·46 CTI			68%
SE	(39·4) (3·9)	(5·2)		

Logistic regression model

$\ln(p/(1-p)) = 2{\cdot}52 + 1{\cdot}68$ umplacrv

re-arranging gives:

p(E horizon + $= \exp(2{\cdot}52 + 1{\cdot}68$ umplancrv$)/(1 + \exp(2{\cdot}52 + 1{\cdot}68$ umplancrv$))$

Analysis of deviance

Model	Deviance	Residual deviance	Df	Pr(Chi)
Null		69·31	49	
umplancrv	29·43	39·88	48	< 0·001

* CTI = compound topographic index.
plancrv = plan curvature.
umplancrv = upslope mean plan curvature.
p = probability that an E horizon is present.

provided the best logistic model fit for the probability of an E horizon occurrence. This indicates that the overall upslope convergent and divergent flow processes may control E horizon development. The next best logistic fit was provided by CTI, which in part, measures some of the same types of landscape processes as upslope mean plan curvature. Figure 3 displays the spatial extension of the logistic model for E horizon presence/absence (figure 3 (*a*)) and the regression model for solum depth (figure 3 (*b*)) for the study catchments.

The advantage of this form of mapping over conventional methods is that individual soil attributes rather than soil types are predicted with a specified accuracy and precision. Assumptions of high covariance between soil attributes, implicit in the mapping of traditional soil types, are avoided. The sampling procedure used here also enables the exploration and identification patterns in the data that may relate to process thresholds in the landscape. Subsequent quantitative delineation of process zones (e.g. zones of net erosion) can be used for land management planning.

5. Conclusions

We began with a provisional pedologic model where CTI was hypothesized to be a strong controlling variable and designed our sampling plan accordingly. The field data supported this assertion and provided evidence of other useful explanatory variables. The identification of plan curvature and upslope mean plan curvature as useful predictors demonstrates a key feature of our methodology. Models are proposed and then tested. During the testing phase, new hypotheses of landscape processes controlling soil distribution are formulated and these may be tested to further improve our capacity for spatial prediction. In conventional surveys, this process is undertaken in the minds of surveyors as they traverse a region and develop mental and sometimes verbal models for spatial prediction.

Our long-term goal is to develop a quantitative and statistical analogue to the conventional method that is explicit, consistent and repeatable. Evidence is not confused with interpretation and models can be communicated in an objective way. At present, a large body of knowledge is trapped within the minds of soil surveyors and is eventually lost. Our procedure meets with Hewitt's (1993) demands for a scientific rather than subjective procedure for developing explicit and quantitative soil-landscape models for spatial prediction. These methods provide a basis for understanding soil-landscape processes and may be integrated with other spatial interpolation techniques such as kriging and splines (Hutchinson and Gessler 1994). Information about scale (Moore *et al.* 1994) and error (Burrough in press) must also be incorporated in an explicit fashion.

Acknowledgments

The authors thank Linda Ashton, John Hutka and Chris Moran for assistance. This study was funded in part by Grant No. NRMS–M218 from the Murray-Darling Basin Commission and by the Water Research Foundation of Australia. Professor Ian Moore passed away before the completion of this paper. He was a friend, mentor and colleague who will be dearly missed. We will endeavour to carry on his spirit and work.

References

ALLEN, F., and STARR, T. B., 1982, *Hierarchy: Perspectives for Ecological Complexity.* (Chicago: University of Chicago Press).

ALLISON, R. J., 1991, Slopes and slope processes. *Progress in Physical Geography*, **4,** 423–437.

BAIZE, D., and GIRARD, M. C., 1992, *Pedological reference base: main soils of Europe, Referentiel pedologique, Principaux sols d'Europe* (Versailles, France: INRA).

BECKETT, P. H. T., and WEBSTER, R., 1971, Soil variability: a review. *Soils and Fertilizers*, **34,** 1–15.

BOUMA, J., 1989, Land qualities in space and time. In *Land qualities in space and time, Proceedings of a symposium organized by the International Society of Soil Science, Wageningen, The Netherlands* (Wageningen: Pudoc), pp. 3–13.

BURROUGH, P. A., 1993, Soil variability: a late 20th century view. *Soil and Fertilizers*, **5,** 529–562.

BURROUGH, P. A., in press, Spatial data quality and error analysis issues: GIS functions and environmental modelling. *Proceedings of the 2nd International Conference on Integrating Environmental Modelling and GIS* (Boulder: GIS World).

BUTLER, B. E., 1964, Can pedology be rationalized: a review of the general study of soils. Presidential address. Publication 3, Australian Society Soil Science, CSIRO Division of Soils, Canberra, Australia.

CHAMBERS, J. M., and HASTIE, T. J., 1992, *Statistical Models in S.* (California: Wadsworth & Brooks).

CLEVELAND, W. S., 1993, *Visualizing Data.* (New Jersey: Hobart Press).

COOKE, R. D., and WEISBERG, S., 1982, *Residuals and Influence in Regression* (London: Chapman & Hall).

FITZPATRICK, E. A., 1993, Introduction: soil horizons. *Catena*, **20,** 361–362.

GESSLER, P. E., MCSWEENEY, K., KIEFER, R., and MORRISON, L., 1989, Analysis of contemporary and historical soil/vegetation/land use patterns in southwest Wisconsin utilizing GIS and remote sensing technologies. *Proceedings of the ASPRS/ACSM Annual Convention* (Falls Church, VA: ASPRS/ACSM), pp. 85–92.

GOODCHILD, M. F., 1986, Spatial autocorrelation. *CATMOG—Concepts and techniques in modern geography* (Norwich: Geo Abstracts).

HEWITT, A. E., 1993, Predictive modelling in soil survey. *Soils and Fertilizers*, **3,** 305–314.

HUTCHINSON, M. F., 1989, A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology*, **106,** 211–232.

HUTCHINSON, M. F., and GESSLER, P. E., 1994, Splines: more than just a smooth interpolator. *Geoderma*, **62,** 45–67.

KACHANOWSKI, R. G., 1988, Processes in soils—from pedon to landscape. In *Scales and Global Change* edited by T. Rosswall (New York: Wiley and Sons), 153–177.

MALANSON, G. P., BUTLER, D. R., and WALSH, S. J., 1990, Chaos theory in physical geography. *Physical Geography*, **11,** 293–304.

MATHERON, G., 1971, *The theory of regionalized variables and its applications. Les Cahiers du centre de morphologie mathematique de Fontainebleu. Ecole Nationale Superieur des Mines de Paris.*

MCBRATNEY, A. B., and WEBSTER, R., 1983, Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. *Journal of Soil Science*, **34,** 137–162.

MCCULLAGH, P., and NELDER, J. A., 1989, Generalized linear models. *Monographs on Statistics and Applied Probability* No. 37. (London: Chapman & Hall).

MCDONALD, R. C., ISBELL, R. F., SPEIGHT, J. G., WALKER, J., and HOPKINS, M. S., 1990, *Australian Soil and Land Survey Field Handbook.* Second Edition. (Sydney: Inkata Press).

MCKENZIE, N. J., 1991, *A strategy for coordinating soil survey and land evaluation in Australia.* Divisional Report 114, CSIRO Division of Soils, Canberra.

MCKENZIE, N. J., and AUSTIN, M. P., 1993, A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. *Geoderma*, **57,** 329–355.

MCSWEENEY, K. M., GESSLER, P. E., SLATER, B., HAMMER, R. D., BELL, J., and PETERSON, G. W., 1994, Towards a new framework for modelling the soil-landscape continuum. In *Factors of Soil Formation: A Fiftieth Anniversary Retrospective*, Special Publication 33 (Madison, WI: Soil Science Society of America), 127–145.

MILNE, G., 1935, Some suggested units of classification and mapping particularly for East African soils. *Soils Research*, **4,** 3.

MOORE, I. D., GRAYSON, R. B., and LADSON, A. R., 1991, Digital terrain modelling: review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5,** 3–30.

MOORE, I. D., GESSLER, P. E., NIELSEN, G. A., and PETERSEN, G. A., 1993, Soil attribute prediction using terrain analysis. *Soil Science America, Journal*, **57,** 443–452.

MOORE, I. D., LEWIS, A., and GALLANT, J. C., 1994, Terrain attributes: estimation methods and scale effects. In *Modelling Change in Environmental Systems*, edited by A. J. Jakeman, M. B. Beck and M. McAleer (London: Wiley), pp. 189–214.

RAYMOND, O. L., 1992, *Geology of the Wagga Wagga sheet, Sheet 8327, Australian Geological Survey Organization, Canberra, Australia.*

SIMONSON, R. W., 1959, Outline of a generalized theory of soil genesis. *Soil Science Society of America Proceedings,* **23,** 152–156.

STATISTICAL SCIENCES, 1992, *S-Plus Programmer's Manual.* Statistical Sciences Corporation, Seattle, Washington.

TUKEY, J. W., 1977, *Exploratory Data Analysis* (Reading: Addison-Wesley).

WAGENET, R. J., BOUMA, J., and HUTSON, J. L., 1994, Modelling water and chemical fluxes as driving forces of pedogenesis. In *Quantitative Modelling of Soil Forming Processes*, Special Publication 39 (Madison, WI: Soil Science Society of America), pp. 17–36.

WEBSTER, R., 1977, *Quantitative and Numerical Methods in Soil Classification and Survey* (Oxford: Clarendon Press).

WEBSTER, R., and OLIVER, M. A., 1990, *Statistical Methods in Soil and Land Resource Survey* (Oxford: Oxford University).